Reality and Morality

Billy Dunaway

University of Missouri–St. Louis

Acknowledgements.

This book is an attempt to give a comprehensive overview of the appeal of what is commonly called "reference magnetism", and natural corollaries in metaphysics and epistemology, for realists in meta-ethics. It is a view about how moral language, and other practical terms, such as 'right', 'wrong', 'ought', etc., get their meanings, but it is more than just a thesis in the philosophy of language. Reference magnetism depends on and exploits a thesis about the *metaphysics* of morality, which will be appealing only to realists: that moral rightness and related properties are objective, metaphysically privileged parts of reality. These action-guiding properties are not mere shadows of how we think or talk about what to do, think, and feel.

While I have been interested in, and written frequently about, realism in meta-ethics, this book is not strictly an adaptation of previously published work. Instead it aims at providing a book-length treatment of the defensibility, and benefits, of reference magnetism for moral realists. I have discussed related issues in previously published work, which most obviously includes "Reference Magnetism as a Solution to the Moral Twin Earth Problem", co-authored with Tristram McPherson, in *Ergo* (2016). I have also written about the metaphysics of realism in "Supervenience Arguments and Normative Non-naturalism", *Philosophy and Phenomenological Research* (2015) , "Expressivism and Normative Metaphysics", *Oxford Studies in Metaethics* (2016), "Realism and Objectivity" in the *Routledge Handbook of Metaethics* (2017), and in my manuscript "The Metaphysical Conception of Realism". At various places in this book I reference work in these papers but do not rehearse the arguments in detail. The aim of this book is to present a defense and articulate a full theory that draws connections between meta-semantics, metaphysics, and epistemology for the realist about morality, which would

not be possible in a series of individual papers. Work on the manuscript was funded by a University of Missouri Research Board Grant during the 2017-2018 academic year.

I have benefitted from discussions and feedback from numerous people when writing various parts of this book. My big-picture influences come from two slightly different directions: my dissertation advisor, Allan Gibbard, is one, and John Hawthorne is the other. Allan helped me to think about disagreement and the distinctive nature of practical language. His books on these subjects repay reading and re-reading, and I am grateful for many conversations with him. John introduced me to the power of Lewis's metaphysics of "natural" (or what I call "elite") properties, and to reference magnetism. I found his thoughts in conversation about how to apply these ideas to morality especially insightful; on many occasions a 30-minute conversation would prove more productive than months of solitary work.

In addition I have benefitted from discussions with a number of people. I am certain to leave some out, but a partial list includes: Charity Anderson, Matt Babb, Ralf Bader, Anne Baril, Anne Margaret Baxley, Matt Benton, Eric Brown, Matthew Chrisman, Jill Delston, Julia Driver, Daniel Fogal, Sam Filby, Steve Finlay, Brian Hedden, Zoë Johnson King, David Manley, David Plunkett, Dani Rabinowitz, Peter Railton, Richard Rowland, Mark Schroeder, Alex Silk, Bart Streumer, and Eric Wiland. In addition I am especially grateful to Matti Eklund, Tristram McPherson, and Mark van Roojen who provided extensive comments on an earlier version of the manuscript for this book. I also wish to thank Peter Momtchiloff at Oxford University Press for his help with bringing this book into existence, and two readers for the press who provided insightful, thorough, and challenging comments.

The most thanks of all go to Rachel, Gabriel, Ellice, and James, for their support

and love.

Table of Contents

Introduction.

Reality favors certain ways of acting over others. This is Matti Eklund's illuminating characterization of a certain kind of view about the nature of morality,[1] which I will call a *realist* view.[2] It is a fact that I morally should pay my taxes and not lie to my friends for no reason. According to the realist, these are facts because of how things are in reality—reality favors paying taxes over not paying, and it favors not lying over lying.

For the realist, reality plays an explanatory role in this description. The most perspicuous explanation of why I morally should pay my takes is something about how the world is—independent of how we think about it, conceptualize it, or want it to be.[3] One job of moral language is to describe these facts.

This book takes realism as its starting point, and does not directly argue for the thesis, or try to defend it. Rather it develops realism as a view not only of the metaphysics of morality, but also its language and epistemology. Many book-length of realism are in the business of assessing realism against its competitors: they aim to defend realism against objections from multiple directions, or to show that it fares better than alternative views.[4] Here I aim instead to partially develop the realist view on its own terms. There is no big picture motivation for doing this. The only claim I will make here is that developing the realist view reveals a range of intrinsically interesting questions for the realist, which can be addressed in a fruitful way. Any support for this claim will have to wait for the main chapters of this book.

[1] Or *ethical*, and more generally, *normative* facts—I return to these distinctions later.
[2] Eklund (2017, 1)
[3] This is a very rough characterization. Dunaway (MS) provides one way of fleshing out this language.
[4] Shafer-Landau (2003) and Enoch (2011) are two important examples.

The framing issue for this project is a related view about language, which it is tempting to think is a corollary of realism. This is not a view only about actual moral language, including the terms 'right', 'should', 'ought' and the like in English. It is a claim about *possible* moral terms, or what possible communities of language users could mean with their moral terms. The alleged corollary holds that *every* possible community who uses expressions that qualify as moral expressions is thereby talking about the same thing. Moral expressions are *stable*, in the sense that differences between uses of moral terms between possible communities do not change what they are talking about. If English speakers, with our term 'right' are talking about the property of moral rightness, then any other community that uses a term that similarly bears on how to act will also be talking about moral rightness.

One motivation for thinking that moral terms are stable in this way is highly picturesque.[5] If reality really favors my paying taxes over freeriding, then there shouldn't be any possible communities that use moral language to talk about something other than moral rightness. Imagine that there were such a community: then, this community might use their moral word 'right' and speak truly by saying 'not paying taxes is morally right'. As 'right' in their language is a moral term, it has bearing on how to act: they not only truly *say* 'not paying taxes is morally right', they then go on to actually not pay. If moral terms mean something different in their language, this community would by their own lights be correct to not pay.

But if reality really does favor paying taxes, then something should have gone wrong with this community. Reality requires their paying taxes just as much as us, and so someone who doesn't pay seems to be missing out on this aspect of reality.

[5] I very loosely follow Eklund (2017, Ch. 1) in presenting this motivation. He uses the term 'ardent realism' for the kind of view that accepts the corollary of stability for normative terms.

Allowing that it is possible for there to be moral terms that don't pick out moral rightness, it appears, entails that such communities aren't mistaken about anything—there are no mistakes in what they believe, or how they act in response to their belief. So a realist should hold that these communities are not possible.

This is one motivation for the thesis that moral terms are stable. It presents itself as a motivation that is specific to realism. I reject it, but will not begin by arguing against it directly. (I return to it in the conclusion.) There are other, related motivations for the similar theses, and I begin with these in Chapter 1. These are cases involving *disagreement*.

Disagreement is a concept we most naturally apply to participants in a conversation, or those who could easily enter into a conversation with one another. When Milton and Bernard disagree about the effects of raising tax rates on the economy (Milton says, 'it will decrease production' and Bernard says 'it will allow for increased spending on resources that the economy needs to grow'), they can have their disagreement by sitting in the same room as each other and asserting denials of the other's position:

Milton: We should not raise taxes because it will harm the economy by decreasing production.

Bernard: No, it will not—raising taxes will produce public resources that a growing economy needs.

But the face-to-face dispute isn't necessary for disagreement. Milton and Bernard disagree with each other even if they never meet. Each might be firmly entrenched in bubbles of like-minded thinkers and never have a chance to deny the other's assertion.

Nonetheless, in virtue of having different opinions about the relationship between the rationality of raising taxes, they disagree.

Philosophers often use a notion of disagreement that extends even further, to speakers for whom it is impossible to have a face-to-face dispute. We can have disagreements with merely *possible* communities. Here morality provides compelling examples.

It is fairly common for English speakers to say things like 'it is morally wrong to lie for personal advancement at the expense of others', and to thereby express a belief about the moral status of lying in certain situations. But it is easy to imagine a possible community which uses a language much like English, where (among other things) they have a term 'morally right' that they apply to different actions. They routinely apply the term to actions that are very different from the ones we say are 'morally right': they say that lying to get ahead is 'morally right', and that helping those in need is 'not morally right'. But all of this is consistent with a further supposition: that when they use 'morally right', they also praise people who do actions which 'morally right' applies to, and when they judge that an act is 'morally not right' (they, like English-speakers, call such acts 'morally wrong'), they blame people who perform such acts. Moral terms in this community have the same *role* in governing behavior and regulating praise and blame; the difference lies in *which* actions this community applies their moral terms to.

There are a number of possible sources of this difference between the self-centered community and us. It is important to focus on the case where this imaginary community uses 'morally right' and its cognates differently simply because of a different moral sensibility. That is, they aren't in a world, and don't imagine themselves to be in a world, where self-centered action somehow works out to be better for everyone.

(Some people think that in a limited economic sphere, selfish competition in a marketplace produces optimal outcomes; in theory a whole community could think that a similar invisible hand operates not only in economics but in every area of personal conduct.) Rather, they are aware at some level of the effects of self-centered conduct on oneself and others; the only difference between them is they, with open eyes, judge the self-centered conduct to be morally praiseworthy, simply because they have a fundamentally different moral outlook than we do.

Even if this community does not exist in the actual world, and so is not one we could ever come to have a face-to-face dispute over the morality of self-centered action with, some philosophers find it natural to say that we disagree with them. I will accept this rough characterization of the relationship, and spend most of Chapter 1 asking how we should characterize it in general terms. The most significant lesson from these disagreements concerns the stability of moral terms. If the self-centered actors are disagreeing with us when they apply their term 'morally right' to blatant cheating for personal gain, then they must be using their term to refer to the same thing that we refer to. They must be talking about moral rightness, the same property we talk about when we use our term 'morally right'. (If they weren't talking about the same thing—if they did not assert that cheating for personal gain is morally permissible—then there would be nothing at issue between us, since this community is not denying what we assert.)

This is a degree of stability for moral terms. Some possible communities that use 'morally right' differently than we do are still referring to the same thing, since moral terms don't refer to something different just because those who use them apply them differently. It also appears to make something distinctive about moral vocabulary. A

community who systematically applied their term 'red' to different wavelengths of light than we do does not disagree with us about whether stop signs are red, even though they will assert the sentence 'stop signs are not red'. They are simply using their term to talk about something other than redness. But moral terms do not work like this: we do not simply interpret the self-centered community as talking about something different. This is an important datum, which needs to be explained.

On its own, this one example does not suggest that the stability for moral terms extends very far. It is one example of a community that uses their moral terms to refer to the same thing that we refer to. It does not show that every possible community uses their moral terms to talk about the same thing. It is of course easy to imagine some variations on this scenario which deliver the same intuitive result: the alternative moral community could be a community of cannibals,[6] or monarchists, or pacifists. Like the self-centered community, they would disagree with us. The differences between them and our own community do not simply make them a community that is talking past us. This is evidence for additional stability for moral terms, and, as the examples pile up, it is tempting to think that *every* possible community which uses some of their words as moral terms will be talking about the same thing as us, and will be capable of having substantive disagreements with us.

But it is false. While a range of communities that use their moral terms differently intuitively do disagree with each other, this is not the case for every community in possession of moral terms. This is the central datum that I will begin with in exploring what the real range of disagreement is, and so what a realist theory should explain. The theory, I will argue, needs to explain why a wide range of possible communities

[6]Cf. Hare (1952)

use moral terms to speak about the same parts of reality. But not every possible community does this: some possible uses of moral vocabulary do have a different subject matter. Generalizing off the usual examples, which I have glossed above, would yield a misleading characterization of the data to be explained. This is the main thesis of Chapters 1 and 2.

An explanation has consequences for the metaphysics, language, and epistemology of morality for the realist. Chapters 3-6 address these points in turn. The bulk of these chapters is concerned with a straightforward *explanation* of why some but not all possible communities have genuine disagreements with moral language. I return in the conclusion to the related issue which is raised by Eklund. Even if we grant that some possible communities use their moral language to talk about something different than what we use our moral language to talk about, threats concerning the objectivity of morality do not arise in the same way in all cases, from the realist point of view. The thesis that moral requirements are an objective part of reality does not, I will claim, require the thesis that every possible community is talking about this part of reality. Taking view of a full range of examples of possible ways to use moral language will make this clear.

One central piece of the positive view I develop is a metaphysical claim. This is a conception of reality in terms of what I call the *metaphysically elite*. What is it for reality to favor certain ways of acting? I assume a proposal that starts with the idea that some properties are metaphysically privileged, or *elite* parts of the world. Chapter 3 develops this metaphysics of elite properties, building off the work of Lewis (1983) and Sider (2012), among others. And it applies the general framework of a metaphysics of elite properties to realism about *morality* in particular; this is the view that properties

like *moral rightness* are themselves elite.[7]

There is a companion thesis to the metaphysical idea that some properties are elite. This is the meta-semantic thesis that elite properties are easy to refer to. Less picturesquely, this is the thesis that among the factors which determine what a term refers to, are considerations involving whether a candidate referent is elite or not. This is sometimes called *reference magnetism.* Reference magnetism in some form has been endorsed by David (Lewis, 1983, 1984), Theodore Sider (2012), and others. Much of the literature on reference magnetism, however, registers significant skepticism about the view.

But it also is a very appealing resource, in light of the data about disagreement from Chapters 1 and 2. In outline, the appeal is as follows: if the property of moral rightness is metaphysically elite then, given reference magnetism, it is in general easier for communities of moral language-users to refer to moral rightness than to other properties. But not *every* such community is in this category: some possible uses of moral language will be better fits with other elite properties. Chapters 3 and 4 address the objections to reference magnetism, both as a general meta-semantic thesis (Chapter 3) and as a thesis about the workings of moral language (Chapter 4). At the end of Chapter 4 I turn to filling in above outline of the appeal of reference magnetism, for a realist.[8]

The realist-friendly explanation of disagreement that I offer at the end of Chapter 4 leaves some questions unanswered. One especially glaring omission is that Chapters 3

[7]This idea has been developed in Dunaway (MS, 2016), Dunaway and McPherson (2016), Suikkanen (2017); see also Fine (2001) and Wedgwood (2007).

[8]The explanation builds on an existing literature: see van Roojen (2006), Edwards (2013), Dunaway and McPherson (2016), and Suikkanen (2017). All of the existing literature focuses on the fact that moral terms appear to be highly stable. They do not, as I do here, focus on the converse fact as well: that the stability of moral terms is not unrestricted.

and 4 will not provide any substantive answers to the question of *which* properties are elite. Chapters 5 and 6 turn to epistemology, and in the process provide some fairly determinate answers to this question. Questions of epistemology are relevant here because epistemology concerns what we can know about moral properties and their metaphysical status as elite properties. Since knowledge is factive—knowing that such-and-such property is elite entails that such-and-such property *is* elite—epistemological claims are directly relevant to filling in some gaps that Chapter 4 leaves open.

Chapter 5 begins with a general epistemology of eliteness. The central idea is that we know what properties are elite on the basis of ordinary first-order investigation. There is no special sui generis methodology for determining which properties are elite. David Lewis (1983) endorses a form of this idea, holding that by discovering that the laws of physics mention mass and charge, we can know on that basis that mass and charge are elite. I adopt a much more expansive view of which properties are elite in Chapter 3; as a consequence the epistemology of eliteness will be more generous than what Lewis allows as well. Thus in Chapter 5 I develop the idea that not only by learning what features in the laws of any theoretical discipline—including not only physics but chemistry, biology, and even ethics—we can come to know that these properties are elite. It is on this basis, I argue, that we should expect there to be elite properties that allow reference magnetism to explain the facts about possible moral disagreements.

Chapter 6 closes by applying the realist theory I have developed throughout the book to a distinct problem of moral disagreement. While my main focus throughout this book is on explaining why we disagree, rather than talk past one another, with our moral language, a separate problem is the fact that we frequently disagree rather than

agree in what we say using moral language. This is often called a failure of *convergence*. Some have argued that the persistence of disagreement, rather than convergence in moral belief, is incompatible with realism. I close by showing how these arguments look given the metaphysics, meta-semantics, and epistemology of realism that I have developed in Chapters 1-5. Since each component of the view is independently motivated, I conclude that arguments against realism from the failure of convergence among users of moral language are not very threatening.

What begins as a question about possible disagreements raises issues in metaphysics (what is it for moral rightness to be elite?), the philosophy of language (how do linguistic expressions, as used by a community, come to refer to a particular part of the world?), and epistemology (how do we know which property moral obligation is, and how do we know which property is elite?). These are large topics on their own. A discussion of the ways in which they interact can get very complicated very quickly.

Since the aim of this book is not to settle issues across all of these subareas, some simplifying assumptions are needed. These are not intended to be entirely neutral assumptions. They are, for the most part, compatible with a wide range of specific views one might take on each topic. Insofar as there is a common theme to these assumptions, they fit with a set of views which are broadly realist in nature: that moral language has the function to describe a part of reality; that being a part of reality involves being metaphysically privileged in some way, and that we can know facts about morality so construed, but can also be mistaken about it.

Here, in more detail, are some of these assumptions. I will not defend them here or later in this book. Instead these are individually natural assumptions for a realist to make, and are not ad hoc or unmotivated. The main purpose of this book is to

explore the consequences of these assumptions, and not to defend them from a neutral perspective. I will mention them here, so that when they appear in the main text, readers will not be tempted to treat them as theses I take myself to be arguing for. In some cases I develop the assumptions in more detail, where such development is needed.

Modal disagreement. The first assumption has already been mentioned above. This is the claim that there is a sense in which two communities can disagree with each other, even if they are not having a face-to-face dispute, and even if they are not capable of such disputes because the communities in question don't even exist in the same possible world. The assumption I am making is that, even in cases of "modal separation" like this, it is possible for one community to disagree with the other about whether it is morally permissible to avoid paying taxes.

There is a further assumption which I will spend significant time explicating in Chapter 1, but will not spend much time defending. This is the assumption that these disagreements are what I call *substantive*. A substantive disagreement is only possible between speakers who use their terms to mean the same thing, and where one denies the literal content of what the other says. There are a number of approaches in meta-ethics which treat these disagreements as non-substantive, in my sense.[9]

Referential semantics. In order to have a substantive disagreement, speakers in the disagreement need to mean the same thing by their terms. In most cases I will say that speakers who mean the same thing with their term 'morally right' and similar expressions do so in virtue of *referring* to the same property.[10] (The near-trivial way

[9] Plunkett and Sundell (2013), Silk (2017), Stevenson (1937), and Gibbard (2003) are some examples.
[10] See Soames (2002) for a theory of meaning along these lines.

to put this point is that the term 'morally right' in English refers to the property of moral rightness. But there may be more informative things to say about what it refers to as well.) Only speakers that refer to the same property with their moral terms are capable of having substantive disagreements about morality.

At some points I will jettison this simplifying assumption, and introduce a complication. This is a theory which is inspired by Kratzer (1977). This view of the meaning doesn't have any analogue of a "referent" for moral terms. But it has some currency with linguists and meta-ethicists working today. It will be instructive to put the main theses of this book in the framework of a Kratzer-inspired theory.

Public language. The central cases of disagreement I will be discussing are couched in terms of *communities* who disagree with each other. The focus on communities is because I assume that languages are public entities, and that the meaning of a term is determined by how a community as a whole uses it.

Theoretical knowledge. The core of realism, as I am using the term here, is a metaphysical claim. But there are related epistemological theses that are often associated with realism. I do not take these theses to be a part of the definition of realism; instead I take them to be additional, although quite natural, assumptions. In particular I will make epistemological assumptions that are frequently associated with realism, including the claims that we can not only know particular moral facts, but in addition can know the true moral theory, and in fact can know this theory when presented in the most metaphysically perspicuous way.

Anti-risk epistemology. When investigating claims about whether we know, or can come to know, the contents of the true moral theory, we need an account of what knowledge

requires in addition to true belief. I will assume that one additional component of knowledge is the absence of *risk* in holding the true belief. Roughly, this amounts to the claim that one has knowledge only if one is not at risk of forming a false belief. Thus I will be asking not only what follows from the claim that we can believe true theoretical claims about morality without at risk of having a false belief, but in addition what this means for our knowledge of facts about the eliteness of certain moral properties.

This book will touch on issues in moral metaphysics, moral language, and moral epistemology. Some framing assumptions, including those listed above, are necessary. I will not address every topic that falls under each heading, since I wish to avoid obscuring the main theses of this book with discussion of issues that are extensively discussed elsewhere. I have little to add to these discussions, so I will rely on simplifying assumptions in order to minimize unoriginal or uninteresting discussion, and to frame the main points of this book more clearly. This will leave many questions unanswered, but it allows for more detailed development of certain aspects of the realist position.

The claims I will develop on behalf of the realist are: *1.* there is an elite property of moral rightness; *2.* there are multiple, highly elite morally relevant properties that are structurally distinct from rightness; *3.* it is possible for communities who use moral language to talk about different properties, and hence to talk past each other and not disagree; *4.* common uses of thought experiments about possible users of moral language are prone to lead to overgeneralizations, and are wrongly used to support the thesis that every possible user of moral language must be talking about the same thing; *5.* the facts about what these possible communities are talking about are well explained by a meta-semantic theory of how language refers to the world that includes reference

magnetism; *6.* reference magnetism is a defensible thesis in meta-semantics for both moral and descriptive language; *7.* we can know what the highly elite moral properties are, since we can read these facts off the deliverances of theoretical reasoning in ethics, and *8.* a theory that accepts these theses should hold that common assumptions about the relationship between realism and convergence are mistaken.

The motivation for these theses begins with some observations about moral disagreement. I save the details for Chapters 1 and 2, but the general starting point is in some respects an odd one for a realist. The scope of moral disagreement is frequently cited as a motivation for *non*-realist theories in metaethics. I begin Chapter 1 with a discussion of the "Moral Twin Earth" cases which Horgan and Timmons, in a series of papers, used in arguments against various versions of realism. This is not an idiosyncratic focus; for example Gibbard (2003) motivates an expressivist theory of normative language by considering the range of possible disagreements between speakers, and a template for this style of argument goes back to Hare (1952).

Insofar as a realist has something to say about disagreement, it appears to be at best the product of a defensive maneuver. The realist can claim that all of the alleged cases of moral disagreement can be explained by tools available to the realist,[11] or claim that *most* such cases can be explained, at least to the extent that disagreement does not constitute a decisive objection to realism, even if shows that there are some unsatisfactory consequences of the view,[12] or deny that the relevant claims about disagreement are significant at all.[13]

What hasn't been done systematically is for the realist to develop a characterization of the range of moral disagreement, in order to claim that it is a theoretical *benefit* of the

[11] van Roojen (2006), Edwards (2013), and Dunaway and McPherson (2016) make claims along these lines.
[12] Perhaps Copp (2000) provides an example of this approach
[13] Dowell (2015)

realist view that it provides a natural explanation of disagreement, so characterized. That is what I do here. The central claim I make in Chapters 1 and 2 is that, while metaethicists have properly emphasized the scope of moral disagreement, they have not to the same extent emphasized that there are some *limits* to the phenomenon. A good account, then, should not only explain why it is that so many possible users of moral language are capable of disagreement with one another; it should also explain why not every user of moral language is in this position.

I argue that the realist can do both, in a natural way. This constitutes an explanatory virtue of the view. I do not argue that no other theory can do the same. This is true in two respects: first, I only develop one version of realism, as an explanation of the relevant facts about disagreement. It may be that other versions of realism can do the same. Second, and more importantly, I do not argue that non-realist theories cannot explain the same facts. But I will note here that it is not altogether obvious that they can explain the limits to disagreement. Typical expressivist (and, more generally, non-cognitivist) explanations appear to generalize too much, entailing that every user of moral language will be capable of substantively disagreeing with others. I do not pursue this objection here; instead I focus on developing the realist view to explain both the scope and limit of disagreement.

The main argument of this book is that there is a version of realism that can explain the relevant disagreement-facts, and moreover the realist needs only some simple and natural theses to deliver the relevant explanation. These include a metaphysical thesis (that some moral properties are elite); a meta-semantic thesis (that elite properties are reference magnets), and an epistemological thesis (that it is possible know which moral properties are elite, because it is possible to have risk-free beliefs about the relevant

facts). It should be a mark in favor of realism, and a significant challenge to to its competitors.

The end result is not a comprehensive realist theory, which addresses every issue that might confront a realist. Instead I take the appeal of the view developed here to lie in two general characteristics of the view. First, it does well in explaining some facts about disagreement which, traditionally, has been an issue which motivates the realist's competitor, the non-cognitivist.[14] Second, the realist explanation I offer of the relevant facts is extremely simple in its essentials. I have developed an application of this kind of framework to other issues elsewhere.[15] This book does not rehearse those arguments.[16] Rather, I aim to show that realism can be developed with these simple theses at its core, as an explanation to some novel, and somewhat surprising, facts about moral disagreement. If this book succeeds in its goals, then further development of realism along the lines I outline here should be warranted.

[14]Perhaps not every version of non-cognitivism is motivated in this way. See Schroeder (2010).
[15]Dunaway (2015, 2017b, 2016, MS)
[16]Some of central meta-semantic claims of the book are developed in Dunaway and McPherson (2016). In this book I build on this point, first by developing the data to be explained in different directions, and then elaborating on the account as it applies to the new explananda.

Chapter 1.

Disagreement, semantics, and meta-semantics.

Disagreements about morality, and about what to do more generally, are easy to come by. This is a striking datum in theorizing about what moral and normative terms mean, and theories of what terms like 'right', 'wrong', 'ought', 'good', and 'bad' mean can take some very different approaches to explaining the datum. But it is something that, in some way or other, needs to be explained.

Before turning to illustrations of this phenomenon, some preliminary points are in order. Moral terms like 'right' can be characterized either by a substantive theory of which acts are morally right, or by a *role* that moral terms characteristically play. Substantive theories of morality are a subject of significant controversy. Some hold that morality requires us to do an act that produces the most overall happiness, regardless of how the happiness is distributed; others hold that morality requires that we do not violate the autonomy of rational agents, regardless of how much happiness the autonomy-violation would produce. Still others hold that right action is connected to virtuous action, or to some notion to human flourishing. Moreover there are differences between theorists (or even proto-theorists found on the street) who fall under each broad heading. We do not need to settle these controversies here. What is at issue in this book is not who is right; what is at issue is the fact that these are controversies in the first place.

Those who endorse different substantive claims about what morality requires will nonetheless use their moral terms with the same *moral role*. That is, to a first approximation, they will characteristically feel guilt when they fail to perform an action they apply 'right' to, blame others who fail to perform 'right' actions, and praise those who

do the actions they apply 'right' to.[17] We can blame, praise, etc. for actions that (fail to) maximize happiness, or we can refrain in cases where autonomy-violations occur. In what follows I will call a a term *moral* if it is used with this characteristic role. One important question is why, and how, disagreements are present between those who use 'right' with a similar moral role, but differ in any number of other ways, including the substantive claims they make about which actions are right.

Some philosophers distinguish between moral uses of 'right', 'ought', etc. and *normative* uses of these terms. Similar questions about disagreement arise here. Some have observed that, even holding fixed that one knows that giving a large sum of money to charity is the morally right thing to do, it makes sense to deliberate about whether to give. Perhaps keeping the money would be prudentially valuable, for reasons that have nothing to do with morality. When one deliberates about whether to do the morally required thing or the prudentially best thing, one deliberates using a *normative* concept. (In addition they might add modifiers like *authoritative* or *categorical* to designate the relevant kind of normativity.[18]) In deciding to give the money to charity, one judges that one ought to give to charity, in the normative sense.[19] But others might use the normative 'ought' to deny this, saying 'it is not the case that one ought to give to charity'. In doing so, they are disagreeing in this circumstance about the normative requirement to give to charity.

[17] Cf. Gibbard (1990)

[18] This distinguishes the normative uses of 'ought' and cognates from other uses that I will not be interested in for the remainder of this book. The rules of chess forbid moving one's rook diagonally, and on this basis we can say things like 'one ought not to move one's rook diagonally'. Someone who lives in a jurisdiction where car owners are required to have their vehicles inspected for compliance with emissions standards truly says 'I ought to take my car for an emissions inspection within the next two years'. In these cases the 'ought' expresses a kind of requirement, since certain actions are required by the rules of chess, or local ordinances. But these are not claims made with the *normative* 'ought', in the sense I am using here. This is because one can coherently accept that the rules of chess prohibit a certain move, but think the all-things-considered thing to do is to violate the rules of chess. And similarly for legal requirements. This is a symptom of the fact that the 'ought's used to state these requirements are not the categorical, authoritative normative 'ought'. I will for brevity omit the modifiers in what follows.

[19] Cf. Gibbard (2003), McPherson (2015).

As with the moral 'ought', there is some similarity between the disputants: the parties disagree using an 'ought' with the same role. Disputants can differ substantially in which actions they claim the normative 'ought' applies to. But they must be relatively similar with respect to the role they use the normative 'ought' with. Since this 'ought' expresses an all-things-considered decision, its role should not allow for further deliberation about what to do. Judging that one ought to donate to charity in this sense closes off deliberation—if one makes the judgment and fails to donate, one has made a mistake. Using an 'ought' with this role does not settle substantive questions of what to do: in a dispute over donations to charity, both parties can agree that *if* one uses the normative 'ought' to say 'one ought to give to charity', then one makes a mistake of some kind if one does not give. This is a part of the shared role in their use of the normative 'ought'. It is compatible with significant differences over which actions the normative 'ought' applies to.

Both moral and normative vocabulary have in common some relationship, in virtue of their role, to how to feel or act. These are *practical* terms. Many of the same issues that arise for the relationship between the distinctive role of practical terms, and the possibility of disagreements that are expressed by statements that include practical terms. Moreover (as I will discuss below) the literature has sometimes focused on these issues purely in the context of moral terms, when the same issue arises for the more general normative vocabulary. So it will be helpful to have a blanket term to discuss these issues in what follows, under the 'practical' heading.

The most compelling examples of disagreements among communities who use their terms very differently are in cases where it is clear that they are using practical vocabulary because of the associated role. That is, the communities in these examples

use their terms 'ought', 'right', etc. with a moral or normative role, and thereby are capable of disagreeing with each other, because there are significant differences in how they use their practical terms in other respects.

This, in broad outline, frames a distinctive feature of moral language, and practical terms more generally. But it does so only in the most general terms, and leaves many details to be filled in. One central issue is the question of when, in particular, terms are used with the same practical role and are on that basis capable of being used to express disagreements. The central datum that much of the contemporary literature has focused on is that these disagreements are quite *extensive:* in fact, the range of possible disagreements seems to be much wider than it is with non-practical terms. There is a second issue that is not explored in anything like the same amount of depth, which is the issue of *how far* the relevant disagreements with practical terms extend. One answer, which is tempting, is that the disagreements extend to *every* use of terms with the same practical role. For example, if someone uses 'right' with a moral role, then they will be capable of disagreeing with every other possible speaker who uses their term with the same moral role. It is not obvious that this is true, but if it is false, no one has said where the limitations to the possible disagreements lie.

This book is dedicated to exploring the foundations of a realist theory of morality and, by extension, a realist theory of normativity. Since the realist holds that there are facts in reality about what we should do, it is natural to interpret speakers who use moral and normative terms as speaking about these facts. But then the existence of disagreement with practical terms places an explanatory burden on the realist: how is it that two speakers manage to speak about the same parts of reality, even when they are making radically different claims about it? Why is it not the case that, when

speakers use their practical terms very differently, they aren't simply speaking about different parts of reality and not disagreeing with one another? These questions are especially pressing within the constraints of the realist view.

But before leveling any challenges to the realist, we would need to know what the explanatory desiderata are. This chapter begins by laying them out in more detail. First, in §1 begins with a survey of some recent literature which explicitly raises explanations of disagreement as a problem for the realist. After laying out the basics of the argument, I turn to elaborating on the central lessons that can be extracted from the argument. Then in §2 I isolate a natural generalization of the explanatory standard, which I formulate under the heading of the **Universal Disagreement** thesis. The rest of the chapter elaborates on what the **Universal Disagreement** thesis would require of the realist, if it were true. I expand on the central notions of moral and normative *role*, and on the notion of disagreement that is at issue in §3. Then in §4 I introduce some complications, and in §5 I distill these lessons into a constraint on the realist's metasemantics, which follows *if* the **Universal Disagreement** thesis is true. This amounts to a claim about the semantic "stability" of practical terms.

1.1 Moral Twin Earth

Perhaps the most influential characterization of moral disagreement is found in the "Moral Twin Earth" scenario from Horgan and Timmons (1991, 1992a,b, 1996). In outline, the point is put in terms of a thought experiment about two separate communities in two separate environments—call these environments 'Earth' and 'Moral Twin Earth'. Earth and Moral Twin Earth are alike in that there are communities which moral vocabulary; each community uses 'wrong' with a moral role. But the

worlds differ in that the communities use their relevant terms so that the most natural interpretation of their speech holds that they are speaking about different properties:

> 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. Suppose, too, that this theory is discoverable through moral inquiry employing coherentist methodology. 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, and so forth (at least those who speak twin English use these terms, whereas those who speak some other twin language use institutions, terms orthographically identical to the corresponding moral terms in the corresponding Earthly language). But on Moral Twin Earth, people's uses 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. (Horgan and Timmons, 1992b, 245)

Horgan and Timmons think it is clear that the communities on Earth and Moral Twin Earth disagree with one another: "here the question about what really is the fundamental right-making property seems to be an open question, and one over which Earthlings and Twin Earthlings disagree".[20] So they are, at least in some cases, *using* 'wrong' differently. These are the cases that constitute a disagreement: the situation is somewhat similar to actual cases where committed consequentialists, who accept a theory along the lines of T^c, appear to be having disagreements with deontologists, who accept a theory like T^d. For example, actual deontologists and consequentialists will disagree about whether it is wrong to steal $5,000,000 from a hedge fund manager, without his consent, in order to provide famine relief to a large population. The

[20]Horgan and Timmons (1992a, 248)

communities on Earth and Moral Twin Earth, according to Horgan and Timmons, will be having the same sort of disagreement when only the deontologists apply 'wrong' to this action.

What the disagreement amounts to, and how far this kind of disagreement extends, is the central question I will be occupied with below. Before turning to these issues, some preliminary points are in order.

First, Horgan and Timmons in surrounding passages make clear that both communities are engaged in *moral* evaluation.[21] But the same point arises if we suppose that they are using normative vocabulary instead.[22] To do this, we only need to imagine that they use 'ought' as a normative term: for instance, they treat someone who applies 'ought' to going to the grocery store as committing the agent who makes the judgment to going to the store, and treating her as incoherent if she fails to go to the store. As in the original Moral Twin Earth case, we can in addition imagine that both communities differ in their use of 'ought' (holding role fixed), treating different actions as all-things-considered required. For instance, we can imagine that on Earth speakers regularly apply 'ought' to happiness-maximizing actions, while on the relevant Twin Earth, speakers regularly refrain from applying 'ought' to actions that violate the autonomous choices of rational agents.

This *Normative Twin Earth* case produces the same verdict as the original Moral Twin Earth case. Each community will say something different about the action of stealing $5,000,000 from a hedge fund manager, without his consent, in order to provide famine relief to a large population. On Earth, speakers will apply their normative 'ought' to the action, saying 'one ought to steal the $5,000,000'. On Normative Twin Earth,

[21] "They use the terms 'good' and 'bad', 'right' and 'wrong', to evaluate actions, persons, and so forth." (Horgan and Timmons, 1992b, 245)
[22] Dunaway and McPherson (2016)

speakers will refrain, saying instead, 'one ought not to steal the $5,000,000'. It appears that there is a genuine disagreement about what to do of faced with the prospect of stealing money in this situation.

The second point to make is that the original Moral Twin Earth case generalizes in other ways. Horgan and Timmons are concerned to make clear that their examples of Earth and Moral Twin Earth will, according to a theory of reference proposed in Boyd (1988), produce a phenomenon of speakers *talking past one another*. One original description of the Moral Twin Earth case emphasizes that there are distinct properties that *cause* speakers to use 'good', 'bad', etc. in the relevant ways. This is important for Boyd's theory; it holds that each community is speaking about the property that is causally related to their usage of moral terms. It appears committed to the conclusion that speakers on Earth and Moral Twin Earth are speaking about different things.[23] This can't be the only difference in the Boyd-centric Moral Twin Earth case, since the relevant communities are stipulated to feel the moralized emotions of blame and guilt in response to different actions, and are claimed to disagree about whether these actions are morally wrong. The case is supposed to be designed so that there is, on Boyd's theory, some actions that fit the following template: a speaker from Earth says 'stealing $5,000,000 from a hedge fund manager is not morally wrong', and speaks truly, because the theft maximizes happiness; whereas a speaker from Moral Twin Earth says 'stealing $5,000,000 from a hedge fund manager is morally wrong', and *also* speaks truly in her own language. This is because the case is designed so that different properties cause the tokenings of 'wrong' in the mouths of each speaker, and so each says true things about distinct properties. There is no disagreement between

[23]This is only a rough characterization: since my aim is not to defend Boyd's theory here, I will not provide the nuances of the theory and engage with the question of whether the full theory can avoid the consequence that the Moral Twin Earth communities are talking past one another.

the Earthlings and Moral Twin Earthlings given Boyd's theory.

The point to make here is that the lesson is not limited just to differences that are semantically significant according to Boyd's theory, i.e., differences in causal relations between properties and practical terms. There are a number of additional dimensions along which the Earthlings and Moral Twin Earthlings could differ, and still be talking about the same thing. Horgan and Timmons have already extended this type of example to other realist theories, including the analytical descriptivism of Jackson and Pettit (1996) and a version of moral functionalism from Brink (1984, 1989). And they go on to argue that the Moral Twin Earth argument can be developed as an argument against any version of moral realism (Horgan and Timmons, 2000, 139-140).

Instead of focusing on these specific variations of the Moral Twin Earth scenario, I will focus on some general ways in which the original description of the case can be changed or supplemented, and still produce a similar intuition that the communities, so described, disagree. Here are a few:

1. Substantive moral/normative theory. Horgan and Timmons provide one example of how, while both using their words with a moral role, two communities might apply their words to different kinds of acts. Their specific case involves a community of consequentialists on Earth, and a community of deontologists on Twin Earth. They do not bother to specify *which* versions of consequentialism and deontology each community accepts. Multiple candidates would do the job: the Earthlings could be Utilitarian happiness-maximizers, or they could be desire-satisfaction-maximizers. The deontologists could accept an absolute prohibition on autonomy-violation, or could instead accept Kantian universalizability constraints on motives. The communities need not even fit into the familiar consequentialist and deontological categories. What is clear

is that there are *many* ways for the communities in a Moral Twin Earth-like scenario to differ in which substantive view about moral obligation best approximates their use of terms like 'right'. But so long as they use 'right' with a moral role, they will appear to disagree.

2. Theoretical reflection. Horgan and Timmons specify that the communities on Earth and Twin Earth would, *if* they were to systematically theorize about moral matters, arrive at distinct theories (T_c and T_d, respectively). The judgment that the communities in question are disagreeing persists whether or not we explicitly stipulate that they are disposed to engage in this high-level reflection, or would arrive at any specific verdict if they did. For instance, take a pair of communities where one says 'self-plagiarism is wrong' and the other says 'self-plagiarism is not wrong'. Assume further that these are not judgments derived from higher-level ethical principles, but rather are direct unreflective reactions from members of each community when presented with a concrete case of self-plagiarism. So long as we specify the cases as involving communities which are sincere in their respective judgments, and where their environments do not differ in ways that could be relevant to the moral permissibility of self-plagiarism, it will still be quite natural to hold that they disagree. Likewise the intuition of disagreement exists if we do stipulate that one, or both, of the communities derives their moral judgment about a particular act from general moral principles.

3. Community-wide unanimity. These simple thought experiments can be run by imagining two communities that are relatively uniform about substantive moral matters. On Earth, there is intra-community agreement on the verdicts of some specific consequentialist theory, and on Twin Earth there is intra-community agreement on the verdicts of

some specific deontological theory.[24] This makes the disagreement between individual speakers in each community especially stark. But it is not necessary. Even if speakers on Earth are not unanimous in using their moral terms 'good', 'ought', and the like as a Utilitarian would, they can still disagree with the Twins. Suppose the community on Earth consists entirely of speakers who use these terms with a moral role, but a minority of speakers on Earth are identical to the Twin Earthlings in using their moral terms in conformity with a deontological theory. An individual speaker on Earth who applies their moral 'ought' to happiness-maximizing acts will still, intuitively, disagree with an individual speaker on Twin Earth who does not.

4. Modal separation. Horgan and Timmons model their Moral Twin Earth thought experiment on a the original "Twin Earth" thought experiment from Putnam (1975). In Putnam's original example, there are two communities on Earth and Twin Earth who use their term 'water' in identical ways: both communities apply their word 'water' to the stuff that they drink in order to stay hydrated, that falls from the clouds in the sky, and flows through (unpolluted) lakes and streams. But Putnam imagines that the environment in which the Earthlings and Twin Earthlings use their language is, though similar at a macro-level, very different at the microscopic level. In particular, the microphysical constitution of the water-like substance of Earth is H_2O. But not on Twin Earth—while the water-like substance there appears, and behaves much like H_2O

[24]Here is how Horgan and Timmons describe the case:

> [T]here is significant, though not perfect, agreement between Earthlings' moral beliefs and Twin Earthlings' twin-moral beliefs. Divergences are reflected in the considered moral (and twin-moral) beliefs with which coherentist methodology begins, and they persist even after the methodology is properly applied by Earthlings and by Twin Earthlings, respectively. These disagreements manifest themselves most prominently in just those cases where consequentialist and deontological ethical theories tend to yield sharply differing prescriptions—cases where consequence-based moral reasoning conflicts with moral reasoning that appeals fundamentally to respect for persons, or to individual rights, or the like. (Horgan and Timmons, 1992b, 246)

does on Earth, the stuff there is *not* H_2O. Instead the microphysical structure on Twin Earth is very different, and the water-like stuff there is constituted by "XYZ", which Putnam tells us is structurally very different from H_2O.

Putnam emphasizes that uses of the term 'water' on Earth and Twin Earth, as he describes them, appear to be talking about *different* things. On Earth 'water' as we use it refers to H_2O, whereas on Twin Earth, the term 'water' as the Twin Earthlings use it refers to XYZ. A speaker from Earth who says 'water is made of H_2O' does not disagree with a speaker from Twin Earth who says 'water is made of XYZ'. Each speaks the truth in their own language.[25] The *Moral* Twin Earth thought experiment is supposed to be significant because we get precisely the opposite verdict when moral terms are involved; speakers on Twin English *do* disagree with English speakers with their moral terms 'right', 'ought', and the like; treating each community as speaking the truth in their own language is not an option. Putnam's original example is clearest when we assume that the speakers from each linguistic community are located in different possible worlds. Since it is very plausible that a world with an XYZ-like microstructure requires different fundamental physical laws, it is arguably not possible for there to be two distinct communities who refer to H_2O and XYZ with their term 'water', but which could come into contact with each other, or otherwise exist in the same time and place. The speakers in the original Twin Earth example are, in the cleanest version of the Putnam thought experiment, separated by modal space.[26]

Modal separation is not necessary for the linguistic communities in the Moral Twin

[25]This expression is borrowed from Hirsch (1997).

[26]It would be coherent to treat Putnam's communities as not separated by modal space, but perhaps only by physical space, if we rejected the assumption that the H_2O and XYZ microstructures require different sets of fundamental physical laws that are not jointly compatible. The point here is that modal separation is a key feature of Putnam's original thought experiment, since it avoids these tendentious metaphysical questions. (Thanks to a reader for raising this issue.)

Earth example. It is perfectly possible for two communities in the same world to use their moral terms differently, in the way Horgan and Timmons describe. They could be in the same world, located on different Earth-like planets where rational creatures with roughly similar capacities and environments have evolved. They could even co-exist on Earth itself: nothing precludes the existence of separate communities who regularly use the word 'right' with a moral role, but with applications that track substantively different moral theories.

We should not take this point too far, however. While Moral Twin Earth thought experiments are not necessarily run with an explicit assumption of modal separation, they cannot be re-run by imagining speakers from the same community who have different dispositions with their moral terms. This is because the explanation for why speakers from the same linguistic community disagree with each other is relatively straightforward and not at all peculiar to moral vocabulary. Since they are part of the same community they intend, in part, to speak the same language as other members of their community. Even if one speaker uses 'right' as a Utilitarian would, and another uses 'right' as a Kantian deontologist would, there is no mystery why they are speaking about the same thing: intentions to share a meaning do the job. The striking phenomenon that the Moral Twin Earth thought experiment points to is that terms used with a moral role appear to mean the same thing even if the users of those terms are entirely separate, and have no intentions to communicate with each other.

Moral Twin Earth thought experiments can be extended or supplemented in a number of ways. I have given four examples. These involve varying the substantive moral theories accepted by the communities in question (whether or not the resulting moral judgments are the product of theoretical reflection), the extent to which the commu-

nities are unanimous on moral matters, and whether the communities are separated by modal boundaries.[27] This frames an important lesson of the Moral Twin Earth thought experiment: it does not simply highlight an idiosyncratic or anomalous feature of practical vocabulary. Instead there appear to be many possible communities who use their terms with a moral role, and all of these communities are capable of disagreeing with one another. This is a significant feature of practical terms, since we could repeat these points using the normative 'ought'. As Horgan and Timmons's contrast with Putnam's original Twin Earth thought experiment shows, ordinary natural kind terms like 'water' do not display similar features. It is a datum that any meta-ethical theory, including a realist theory, should try to explain. But this is just an informal characterization of the explanatory target. Before turning to an explanation, we should characterize it in more detail.

1.2 Universal disagreement

1.2.1 A preliminary characterization

The Moral Twin Earth thought experiment—and its cousin, the Normative Twin Earth thought experiment—tells us something about the relationship between different possible communities of users of practical language. In particular, it tells us that, among the communities that use a term with a moral or normative role, they *disagree* with other possible communities who use their terms with the same role, but differ in other respects. This raises the question: *how far* do the disagreements extend across modal space? There are many possible linguistic communities, who use their language

[27] The "missionaries and cannibals" example in Hare (1952) gestures at the same phenomenon. I will not discuss this example in detail since, among other things, it makes use of communities who have causal contact with one another and so introduces additional constraints on interpretation that are not present with modally separated communities.

in different ways. In the previous section we gave a piecemeal description of some possible communities that disagree with each other with practical terms; now we can ask what a general characterization of the disagreement-relations looks like. [28]

It is very natural to extrapolate from these examples to the claim that *any* two possible communities that use a term with the same practical role will disagree with one another. That is: shared practical role is sufficient for disagreement. Any other differences between two possible communities, so long as they bother use 'right', 'ought', or some other term with the same moral or normative role will produce a disagreement between them (and will not result in a situation where they are talking past one another).

I will call this the **Universal Disagreement** thesis:

Universal Disagreement For any two possible worlds w and w^* and linguistic communities c and c^*, if c is in w and c^* is in w^* and c and c^* both use a term with the same practical role R, then c and c^* are thereby capable of disagreeing.

More colloquially, **Universal Disagreement** says that if two communities use 'right' with the same moral role, then—no matter what other differences there are between them—these communities are capable of disagreeing, in the way the original Moral Twin Earth communities disagree. An analogous point holds for the normative 'ought'.

The **Universal Disagreement** thesis will be refined in a number of ways in what follows, and I will ultimately reject it in Chapter 2. Before turning to these projects,

[28] I have granted that there are also possible communities that are in the same possible world, differ from each other in the relevant respects, and thereby disagree. Given a plausible principle of recombination (Lewis, 1986, 88), these communities that have intra-world disagreements will also have identical counterparts that are parties to an inter-world disagreement. So nothing is lost in the modal formulation of the question that I give here.

it is worth elaborating on why it is at least somewhat natural to accept **Universal Disagreement**.

1.2.2 Motivations

Why accept **Universal Disagreement**? One simple motivation is a generalization off of the Moral Twin Earth case and variants. These cases support a thesis which entails that disagreements with practical terms are fairly common across modal space. But **Universal Disagreement** is a stronger thesis, so we should look for additional motivations.

Robbie Williams (2018, forthcoming) assumes a thesis which, given assumptions I will develop later in this chapter, entails **Universal Disagreement**. Normative terms are, according to Williams, *referentially stable*. Reference is a semantic property of a term: what a term means is explained (in part) by what it refers to.[29] Stability is a modal feature of reference: expressions are referentially stable to the extent that they refer to the same thing in different possible environments.

For Williams, practical terms are stables across their role. This means that, if R is the role that characterizes the distinctive practical role of moral terms, then there is a property P that *every* R-playing term refers to:

> Stability says: necessarily, if an agent has a concept W that plays role R, then W denotes property P. (Williams, 2018, 42)

What is the role R? In outline, the role for a moral term is given by its distinctive connections with motivation, blame, regret, and praise. For instance Gibbard (1990,

[29]Some caveats: 1. singular terms refer to objects; for the most part I will treat practical terms ('right', 'ought', etc.) as predicates, which refer to properties; 2. at times I will make an exception to the foregoing, as in parts I introduce the complicating assumption that 'ought' is an operator, with the semantics roughly as given by Kratzer (1977); 3. for the same of simplicity, for the most part I ignore aspects to meaning besides reference.

43) says that "what a person does is morally wrong if and only if it is rational for him to feel guilty for having done it, and for others to be angry at him for having done it".[30] This is one candidate characterization of what we can call *wrongness-role*. One uses 'wrong' with the wrongness-role when one feels guilty for doing actions one applies 'wrong' to, and when one blames others for performing these actions. There are related roles for 'morally right' and other moral terms.[31]

The upshot of stability, in Williams's sense, is that the capacity for disagreement across all communities that use a term with the wrongness-role. Given stability, they will use their term ('wrong', or some synonym) to refer to the same property, namely wrongness. Differences in use of this term will be capable of producing disagreements: one community, using 'wrong' with the wrongness-role, might say that a joke that causes no harm but brings joy and mirth to many is not wrong, on the grounds that it maximizes happiness. Another community, concerned with avoiding even harmless slights to an individual's honor, says that the joke is wrong. This is a disagreement, if they are both referring to the same property, since each disagrees about whether a specific act—the harmless joke-telling—instantiates the property of wrongness or not. The same goes for every difference that is compatible with a shared moral role.

Eklund (2017) raises a related issue. The central claim concerns the undesirable consequences of a *failure* of stability: if different communities could use normative terms to talk about different things—and hence talk past one another—a certain kind of *symmetry argument* would be in the offing. Eklund puts the point as follows (where the "Bad Guy" is a member of a possible community that hypothetically uses their normative terms to talk about something different than what we talk about):

[30] *Cf.* Williams (2018, fn. 2); see also Darwall (2006).
[31] Though see Merli (2008) and Björnsson and McPherson (2014) for worries that this thought can be adequately developed. Eklund (2017, 10) admits that the notion is somewhat obscure.

> We can still say that Bad Guy doesn't do what he all-things-considered ought to do or has reason to do. But using his language, Bad Guy can say the corresponding things about us. Using his counterpart of "wrong"—the word in his vocabulary that has the role for him that "wrong" has for us—he can say that we do "wrong" things. And he is as correct in his verdict about us as we are in our verdict about him. The same would go for all other normative vocabulary [...] Despite all the realist trappings that our normative language is supposed to have, there may still for all that has been said be *parity* between us and Bad Guy that the ardent realist would want to avoid. For all that has been said, Bad Guy is not objectively mistaken about anything; he just does not employ our notion of reason or our notion of what ought to be done but instead employs alternative normative notions. (Eklund, 2017, 5)

There is symmetry between Bad Guy and us on the assumption that stability fails: whenever we make (true) claims about how Bad Guy is mistaken, he can make analogous claims about how *we* are mistaken. His claims are true in his language—ex hypothesi, he is speaking a language where his 'wrong' refers to something different.

There are several important caveats concerning Eklund's claims about Bad Guy. First, he does not say that *every* theorist faces pressure to deny that Bad Guy speaks truly. But he does say that a realist (Eklund's label for the view is Ardent Realism) should be uncomfortable with this, since the realist will want to claim that there are facts in reality about what we should do, and these same facts apply to Bad Guy as well. Second, he does not say that it is required *by definition* for a realist (even of the Ardent variety) to endorse stability. I won't go in for a detailed discussion of concessive approaches to Bad Guy, but for now we can note that Eklund finds these stability-denying options unsatisfying.[32]

If a symmetry argument puts pressure on us to concede that Bad Guy is referring to the same thing we are referring to with our 'wrong', then plausibly there are other possible scenarios that suggest, for similar reasons, that the speakers in them use their

[32]See McPherson (2018) for more discussion.

practical language to refer to the same thing we are referring to. Whether the original Bad Guy case can be extended so far as to support Williams's stability thesis—and hence the **Universal Disagreement** thesis—is an open question, which Eklund does not take a stand on.

Finally, recall the optimism from (Horgan and Timmons, 2000, 139-140) that there is a recipe for a Moral Twin Earth argument against every version of realism that might be put forward. The original Moral Twin Earth argument works, in part, because it seems clear that the possible communities are disagreeing with each other. In order for the recipe to be truly general, we need the assumption that the communities that are cooked up by the recipe to disagree with each other. One natural basis for this assumption is that the **Universal Disagreement** thesis is true.

To sum up: the original Moral Twin Earth thought experiment does not, on its own, entail **Universal Disagreement**. But it, plus other variations, does suggest that there are many possible communities across modal space that can disagree with each other. One natural and tempting generalization of this phenomenon is the **Universal Disagreement** thesis—all possible communities that uses a term with the same practical role are capable of disagreement. Moreover, the generalization **Universal Disagreement** lurks as an open possibility in recent discussions of realism. Not all of these motivations straightforwardly *entail* **Universal Disagreement** (Williams is the exception here). Regardless, they emphasize the *extent* of stability and disagreement with practical terms, but not any potential limits.

Any explanatory project, including that of the realist, will need to know where the limits are, before developing a meta-semantic theory that explains the disagreement-facts, including those in the original Moral Twin Earth case. If **Universal Disagreement**

is true, then the theory should explain it, or else incur the costs of failing to do so. If it is not true, then a good theory is tasked not only with explaining the extent of moral and normative disagreement across modal space, but also the cases where communities use practical terms and fail to be capable of disagreeing.

1.3 Refinements

In Chapter 2 I will argue that **Universal Disagreement** is false. The remainder of this chapter aims to characterize what we should take the **Universal Disagreement** thesis to be. The explicit formulation of the thesis is as follows:

Universal Disagreement For any two possible worlds w and w^* and linguistic communities c and c^*, if c is in w and c^* is in w^* and c and c^* both use a term with the same practical role R, then c and c^* are thereby capable of disagreeing.

The thesis, as stated, raises the following questions: *1.* which *communities* are at issue, *2.* what it takes for a role to be *moral* or *normative*, *3.* what it takes for a community to use a term with a moral or normative *role*, and *4.*. what *disagreement* amounts to. I take these questions in turn.

1.3.1 Possible communities

Meaning is determined in part by how an entire linguistic community uses their terms. There are two reasons for this emphasis on the community-wide aspect to meaning-determination: first, an individual's pattern of usage (along with their dispositions to use) will be too sparse and uninformed to provide a supervenience base for determinate semantic facts.[33] Second, individual speakers intend to mean what

[33] I borrow the terminology from Manley (2009).

other speakers in their community mean—without intentions to speak the same language, communication about a shared subject-matter would be difficult. Even if it is possible for some speakers to have a (perhaps wildly indeterminate) language without intentions to mean what other speakers mean, such speakers will be quite rare across modal space given the communicative function of language.[34]

There are many possible ways for a community to use the items that make up a public language. There are no trivial connections between the similarities and differences in use between these possible communities, and the semantic facts concerning whether these communities mean the same thing as each other. The **Universal Disagreement** says that *all* possible linguistic communities meet the following condition: if any two of them are using a term with the same moral or normative role, they are thereby capable of having a disagreement with that term.[35]

The modal implications of **Universal Disagreement** can inspire skepticism about how we could know that such a thesis is true. It rests, in particular, on the claim that we can know that particular possible communities—for instance those described in Horgan and Timmons's thought experiment—are talking about the same thing, and are not talking past one another. Dowell (2015) provides one kind of skeptical take on this claim. Judgments of substantive disagreement are judgments to the effect that speakers from each community are using a practical term such as 'ought' to speak about the same thing—call this property *obligation*. A judgment that a speaker uses

[34]To avoid interpretive difficulties, I make no claims about the similarities or differences between these points and the "Private Language Argument" in Wittgenstein (1953). The same points go for mental content and representation, which I assume in many cases will be dependent on the meaning of items in a public language.

[35]Note that 'use' with the relevant role is not restricted to they needn't use the same strings of phonemes or lexicographic inscriptions. English-speakers use terms like 'ought', 'right, etc. to express moral and normative judgments, but **Universal Disagreement** is not restricted to communities that use the same strings. A French-speaker who uses her term 'devrait' with the same normative role can stand in disagreement-relations with an English-speaker's normative use of 'ought'.

their term 'ought' to refer to obligation is a semantic judgment since it is a judgment about what a term refers to. Moreover, Dowell claims, semantic judgments of this kinds are known in virtue of our semantic competence with a language. But, the argument goes, semantic competence can't explain how we know what a separate community in a Moral Twin Earth scenario is talking about. Ex hypothesi, such a community is not necessarily speaking our language—that is, they are not necessarily speaking a language where all of their terms that are orthographically and phonetically similar to ours have the same referents.[36] There is no reason to expect that, on the basis of our semantic competence, we should be able to know what this possible community is referring to.[37]

This intermediate conclusion, as stated, is surely correct. As Dowell explains, any good theory of competence with our own language do not explain how we know what speakers using other languages, that are not our own, are referring to.

But the conclusion of this argument does little to support skepticism about semantic judgments concerning possible linguistic communities. In order to secure this conclusion, we would need to accept that it is our semantic competence with our own language that is the only available basis for our judgments of disagreement. It is difficult to see why we should accept this claim.[38]

It is sometimes reasonable to reject the demand that we provide an explanation of how we come to possess the knowledge we have. Knowledge is compatible with ignorance of how we come to acquire the knowledge. Consider a community of scien-

[36]It is true that, as these thought experiments are set up, it is supposed to *follow* from the description that the community refers to the same thing as we do. But this isn't built into the description of the case itself; rather, it is supposed to be something we know on the basis of the description. Dowell is calling into question this aspect of the thought experiment.
[37]Dowell (2015, 11)
[38]See also Eklund (2017, Ch. 5).

tific neophytes: they have, among other things, no knowledge of how vision or light work, and do not give much thought at all to the way in which their sensory faculties work. Nevertheless they typically believe that deliverances of their visual system, and typically form true beliefs on this basis. It is very natural to say that have some basic knowledge of their perceptual environment, even if they are not in a position to explain how they have this knowledge, if asked. Similarly: it seems clear, absent additional reasons for doubt, that the Moral Twin Earth community is talking about obligation (or moral rightness, etc.) just as we are. Dowell has pointed out that we don't know this on the basis of our competence with the English language—the Twins needn't be English-speakers, and even if they are, it is not our competence with English that underwrites this judgment.

I will not try to offer an alternative theory to offer in its place. But I will part ways with Dowell since I will be assuming that we can know semantic facts about possible communities who are (potentially) using their words to mean something different than what we mean with ours. This methodological commitment is shared by motivations for **Universal Disagreement**, and is one I will rely on when rejecting the thesis in Chapter 2.

1.3.2 Morality and normativity

Universal Disagreement is supposed to capture something special about practical language, which distinguishes it from descriptive language. I will return to this point in discussing the notion of semantic stability below. But this raises a prior question: what makes the terms 'right', 'ought', and the like in English count as practical terms in the first place? And what would it take for a term in another language to count as

a practical term?

The answer, in outline, is that they are conventionally used with a practical role—that is, with systematic connections to action and emotion. Here I will provide some suggestive accounts of the characteristic roles that distinguish different types of practical roles. These are not intended to be definitive. Rather, these examples will provide a concrete picture of the issues surrounding **Universal Disagreement**. Slight modifications to the accounts I present here will not require substantial modifications to the main points about **Universal Disagreement** that I make in the next chapter.

Begin with moral terms. In outlining Williams's commitment to **Universal Disagreement**, I have already roughly characterized is distinctive of moral evaluation: in the case of moral disapproval, this involves feelings of guilt when one performs an act one judges to be 'wrong', and feelings of blame to others. This involves characteristic moral emotions: blame, when directed at others who perform "wrong" actions, and guilt, when directed at oneself. We can call this the *moral wrongness role:*

> A term t is used by a community with the moral wrongness role just in case speakers treat it as appropriate to blame other agents who perform actions that the community applies t to, and appropriate to feel guilt when they perform actions the community applies t to.[39]

Conversely, we can say that 'morally right' is used with the *moral rightness role*. When speaking about moral terms generally, including those with both positive a negative valence, I will say these terms are used with a *moral role*.

[39]Cf. Gibbard (1990, 42), **?**, Ch. 4WilliamsNatRep. We might consider some additions to the characterization of the moral wrongness role: perhaps it also involves the fact that moral terms play a distinctive role in a kind of impartial, social regulation of action. Darwall (2006) and Foot (1958/9, 92) suggest upstream constraints on which types of action genuinely moral terms can be applied to.

Speakers can use 'wrong' with the moral wrongness role while endorsing a range of substantive opinions about what moral wrongness consists in. This is a central feature of the original Moral Twin Earth case. Embezzling money from one's employer might have extremely good consequences, overall: Earthling consequentialists in Horgan and Timmons's example will not apply 'wrong' to such an action, and so (since they use 'wrong' with the moral wrongness role) will not blame the embezzler, or expect them to feel guilty. By contrast the deontologist Twin Earthlings will plausibly apply 'wrong' to a very similar act of embezzling, since it involves deception and a use of property that conflicts what the owner would wish to happen with the property. The Twin Earthlings also use 'wrong' with the moral wrongness role, and so they will blame the embezzler, and expect her to feel guilty.

The distinction between moral and normative language amounts to a difference in role. Morality bears on a certain type of morally relevant action and response; normativity is the most generic type of obligation that bears on action. Some have held that it makes sense to deliberate about whether to do what morality requires. This is explained by saying that one can wonder whether one should *in the normative sense* do what one is required *in the moral sense* to do. The normative 'ought' and cognates are distinguished in virtue of the fact that they are used with a *normative* role.

This insight is developed in Gibbard (2003, 2013). Suppose I acknowledge that I morally ought to give a substantial amount of money to charity. I might still wonder if I always need to do what morality requires, and so wonder whether I should give the money. Here I deploy the term 'should' with a normative role. Gibbard elaborates on this in the following way: if I decide that I should give the money, then it rules out not giving. Thinking that I should give the money, in the normative sense, and not doing

so, is inconsistent:

> To believe that one ought to do something is in some way to favor or settle on that action [...] For the special tie of oughts to action, the metatheory of this book says what tie is invariant. The invariant tie is conceptual, a matter of entailments and consistency. What one ought to do settles what to do in the following sense: *It is conceptually inconsistent to believe I ought right now to do a thing and to act otherwise.* This helps characterize the sense of 'ought' that I have been regimenting. I can't consistently believe I ought right now to leave my burning building and decide to stay. (Gibbard, 2013, 224)

We can summarize Gibbard's view on the normative role of 'ought' as follows:

Thinking that one ought to ϕ and not ϕ-ing is inconsistent.

Call this the *Gibbard role*.[40] A central theme of Gibbard's work is that the Gibbard role exhausts the conceptual demands of 'ought'. One can use the term, consistently and without confusion, in many different ways. Just as users of the moral 'wrong' on Earth have counterparts on Twin Earth that use the term with the moral wrongness role, but disagree over substantive questions concerning which actions are morally wrong, similarly two possible communities might use a term 'ought' with the Gibbard role and have substantive disagreements about what ought to be done.[41]

Practical terms are, in what follows, terms that are used either with a moral or normative (i.e., Gibbard) role. **Universal Disagreement** says that any two possible communities that use 'wrong' with the moral role will be capable of disagreeing with each other; similarly any two communities that use 'ought' with the Gibbard role are capable of disagreeing. This is a natural explanation of the extensive disagreements across modal space, which are pointed to by Moral Twin Earth thought experiments.

[40] For alternative characterizations of normative notions see Wedgwood (2001, 2007).
[41] Cf. This is elaborated on in Dunaway and McPherson (2016).

1.3.3 Roles

We have said what is characteristic of moral and normative roles. But we haven't said anything about what it is for a term to be used with a role (whether moral/normative or not). Some clarity on this matter will be important for understanding the commitments of the **Universal Disagreement** thesis.

Start with some obvious claims: a community that used a term 'ought' whenever they are in the presence of red objects, rarely feels any motivation to do anything in particular when tokening the term 'ought', and never expects others to do anything when they deploy the term, does not use 'ought' as a normative term. There is no distinctive commitment to action that they expect to be accompanied by tokenings of 'ought'. In their language, 'ought' is not a normative term.

Even though in this community the word 'ought' is not a normative term, it might on some particular occasions bear the distinctive features of normativity. For instance take Bob, a member of this community who is particularly attracted to red things. While using 'ought' non-deviantly (for his community) Bob will also feel a motivation to possess the things he applies 'ought' to. He does this because of an additional psychological feature that is specific to him, which is an attraction to red things. While this is an aspect of usage of the term in Bob's community, it does not elevate 'ought' to the status of a normative term. There are a number of features that contribute to the distinction between 'ought' in Bob's language and a term that has a a genuine normative role in a public language. These include:

1. Communal agreement. Bob has a particular motivational profile which moves him to act when he tokens his term 'ought'. This is not a community-wide feature how-

ever; other speakers in Bob's community are not as concerned with red things. The particulars of Bob's usage are not expected by others in his community.

2. Psychological contingency. It is possible for Bob himself not to have this profile: his use of 'ought', depending on the particulars, would not have changed in meaning if he hadn't had an interest in red things. Moreover these are not distant possibilities; there are nearby worlds where Bob's interest in red things does not exist.

3. Conceptual connections. Bob doesn't encode the connection between the things he applies 'ought' to and motivation in general form. That is, he doesn't systematically represent applications of 'ought' as entailing motivation. This is unlike our use of 'bachelor' in English which is systematically connected to tokenings of 'unmarried', and unlike the official characterization of the Gibbard role. The connection isn't just a reliable co-application, where we are disposed to apply 'unmarried' to every individual we apply 'bachelor' to. 'Bachelor' is distinctive because it is not only reliably applied alongside 'unmarried' by English speakers; in addition they treat each other as making a mistake if they deny that 'unmarried' applies to something that 'bachelor' applies to. Using 'ought' with the Gibbard role requires something similar, which is missing in Bob's case: no one in Bob's linguistic community will treat Bob as making a mistake if he doesn't attach motivation to his deployment of 'ought'.

The lack of these features is a good indicator that Bob's use of 'ought' does not qualify it as a normative term. This is not a full theory of what separates role from other aspects of use. But it is suggestive: aspects of use will not qualify as a role if there fails to be a significant amount of communal agreement, if they are connected to

highly contingent psychological features, and if they are not encoded as fairly general conceptual connections by the relevant linguistic community.

Nonetheless we can't understand roles to involve instantiation of these features to the highest degree. Not every person is invariably motivated by their normative judgments; even for a normative term the motivational connection does not have to be invariant across the entire linguistic community.[42] Calling an aspect of usage a *role* distinguishes it as fairly systematic and robust, but the distinguishing factors come in degrees. This raises an important point, which is central to the positive theses of this book. Roles are one central and particularly important aspect of use of a term, which play a role in determining what it refers to. But they are not the *only* meaning-determining feature of a linguistic community. Saying that a community uses a term with a moral or normative role settles some relevant meaning-determining properties, but it does not settle all of them.

Since there are aspects to the use of practical terms that go beyond what makes such terms moral or normative, we should distinguish several versions of **Universal Disagreement**. These versions of the thesis take different stances on the contribution of shared practical role to disagreement. Roughly, we can divide these stances into that make claims about possible communities that share a practical role while potentially differing in any other respect, and those that make claims about a subset of these possible communities.

Begin with a statement of **Universal Disagreement** in simplified form:

Universal Disagreement Any two communities that use a term with the same practical role are thereby capable of disagreeing with each other with that term.

[42]See, for example, discussion in Smith (1994), Shafer-Landau (2003), and Wedgwood (2007).

One precisification of this thesis is that communities will disagree when the practical role of 'ought' or a similar term is the *only* role 'ought' is used with. Another thesis is that such communities will disagree no matter what other roles the communities in question might associate with their 'ought'. These are the **Cautious Universal Disagreement** and **Ambitious Universal Disagreement** theses, respectively:

Cautious Universal Disagreement Any two communities that use a term with the same practical role *and no additional roles* are thereby capable of disagreeing with each other with that term.

Ambitious Universal Disagreement Any two communities that use a term with the same practical role *and possibly differ in which additional roles they use the term with* are thereby capable of disagreeing with each other with that term.

The difference between **Cautious Universal Disagreement** and **Ambitious Universal Disagreement** (***Cautious*** and ***Ambitious***, for short) is one of scope. **Cautious** makes a claim about the possibility of disagreements between communities that use a term with the same practical role, but no other roles. The original Moral Twin Earth case is one which Cautious says (correctly) will be a case where there is disagreement: each community uses 'right' with a moral role, and no other roles.[43] A complete picture of practical terms should not content itself with explaining **Cautious** only. Instead we should also ask whether **Ambitious** is true as well—that is, whether any two possible communities that use the same practical term are capable of disagreeing

[43]Even if the communities are unanimous in using their term 'right' in accordance with a specific deontological or consequentialist theory, these aspects of their respective uses would not amount to an additional role, in the sense I sketched above. It is no part of the original Moral Twin Earth case that these communities encode the claims of the deontological or consequentialist theory as a conceptual connection that governs their use of 'right': neither community is disposed to treat the other as simply confused, or incompetent. Below I discuss cases where these substantive claims are conceptually encoded.

with each other, no matter what additional differences there are between them—and if **Ambitious** is not true, we should explain why not.

The difference between **Cautious** and **Ambitious** captures important claims about the meta-semantic significance of a practical role. If **Cautious** is true, then practical role plays some part in determining what a community means by their practical terms. It ensures, in the absence of other roles, that two possible communities are capable of disagreeing with each other. **Ambitious**, if true, requires a stronger semantic contribution from practical role: practical role would, given **Ambitious**, render all other features irrelevant. Once we know that two communities use their term with the same practical role, we would be able to know that they are capable of disagreeing with each other, regardless of what other differences there might be between them.

When asking how far across modal space possible disagreements with practical terms extend, **Ambitious** is a live option—and is not one that is ruled out by existing discussions of the scope of such disagreements. Whether and how it fails is an important data point for any rigorous development of the meta-semantics of practical terms.

While I will reserve the main discussion of **Ambitious** for Chapter 2, there are qualifications that are in order.

Ambitious in one respect is clearly too strong. There is a clear sense in which a community might use their term 'ought' as a normative term, but include in addition some definitional stipulations that make their term fail to be suitable for expressing disagreements with other possible users of normative language. For example, consider a community that uses 'ought' with the Gibbard role, but in addition treats anyone who fails to use 'ought' in accordance with a specific consequentialist theory as not

merely wrong, but conceptually confused. I will call this a *definitional role*. It is certainly possible for a community to use 'ought' with both the Gibbard role and this additional definitional role.

This definitional role for 'ought' would be analogous to a community that used 'bachelor' with its familiar meaning, which makes 'all bachelors are unmarried males' true, but in addition treated it as a definitional truth that a specific individual—call him 'Cyrus'—is a bachelor. (By stipulation, they treat it as definitional of 'bachelor' that it applies to Cyrus; not that it applies to a person with the general properties that Cyrus happens to have.) It is not straightforward that this community would be disagreeing with us over whether Cyrus is a bachelor when they say 'Cyrus is a bachelor' and an English-speaker says 'Cyrus is not a bachelor'. Since it is supposed to be a matter of definition in their language that 'bachelor' applies to Cyrus, but not in English, intuitions of disagreement are not strong. The same goes for a community that uses a normative 'ought' with an additional definitional role which requires that it applies to certain actions that are obligatory according to a specific consequentialist theory.

These additional definitional roles risk giving rise to what Eklund (2017, 14) calls a *defective predicate*. If in fact the relevant consequentialist theory is not the correct theory of normative obligation—or if in fact Cyrus is not a bachelor—then there is a sense in which using the expression is objectionable.[44] The objectionable nature of the additional role for 'bachelor' is clear; someone who used the term with both of its associated definitional roles would say something false. This is the kind of defectiveness that Eklund (2002) claims is present in languages that are inconsistent,

[44]Even if Cyrus is unmarried, the additional definitional role for 'bachelor' is still objectionable in this sense, because his marital status is contingent.

on his gloss. The defectiveness in the normative term is less straightforward: someone who used it with its additional definitional role would, if they continued to treat it as a normative term, end up doing things that they are obligated not to do. But they would not thereby have contradictory *beliefs*, simply in virtue of their competence with 'ought' in a language where it has a definitional role.

This shows there are some restrictions on an absolutely unrestrained version of **Ambitious**, since a normative 'ought' with a definitional role will refer to whatever satisfies the associated definition. But there are many possible communities that do not use their practical terms with a definitional (and potentially defectiveness-introducing) role of this kind. It is not the case that simply any role, when used in conjunction with a bare moral or normative role, gives rise to a defect of this kind. So it is a worthwhile question to ask whether **Ambitious** is true, excepting cases where one of the possible communities in questions accepts additional definitional constraints which directly determine what the practical term in question applies to. Call practical terms that are free of such definitional constraints *primitive*. The generalization in our sights is then the following:

Primitive Ambitious Universal Disagreement Any two communities that use a primitive term with the same practical role are thereby capable of disagreeing with each other with that term.

1.3.4 Disagreement

Primitive Ambitious Universal Disagreement is a claim about *disagreement*. The term 'disagreement' is often used informally with an intuitive meaning. But **Primitive Ambitious Universal Disagreement**, and the specific cases that motivate it, require

something more precise. One (and perhaps the ordinary) notion of disagreement is an inter-personal notion, where the participants are in communicative contact with each other, and attempting to convince, or at least have some influence on, the other party to the dispute.[45] But **Primitive Ambitious Universal Disagreement** does not make claims about these kinds of disagreement, where participants can be assumed antecedently to be speaking the same language. Instead it expands the notion of disagreement to a relation between entirely separate linguistic communities, each of which instantiates meaning-determining properties independently of the other.

The disagreements at issue are *inter*-community disagreements. In an inter-community disagreement the speakers need not be aware of each other, and are possibly modally separated. The facts about disagreement cannot be explained by intentions to communicate with one another, in these cases. Whether two communities do disagree with each other is determined by semantic features that are intrinsic to each community and their environment. If these features antecedently make it the case that they are talking about the same thing—moral rightness, normative obligation, and the like—rather than talking past one another, then they are capable of disagreeing.

What is inter-community disagreement? One simple heuristic is: two communities c and c^* disagree when members of c accept a claim that is logically incompatible with a claim that members of c^* accept. But this heuristic is too simple for disagreements between communities who are modally separated.[46] For instance in Putnam's original Twin Earth case, someone on Earth might believe that the stuff in the Pacific Ocean is H_2O. On Twin Earth, someone in a nearly identical situation at the macroscopic

[45] Brink (2001, 171). This is the kind of disagreement that Stevenson (1937) uses to illustrate his notion of "disagreement in attitude": face-to-face disagreements involving speakers of the same language. These speakers share communicative intentions and practical interests.
[46] See related discussion in Cappelen and Hawthorne (2009).

level believes that the stuff in the Pacific Ocean is not H_2O (they correctly believe that it is XYZ). These contents are logically inconsistent with each other. But there is no natural sense in which the speakers on Earth and Twin Earth disagree—as Horgan and Timmons note, Putnam's original Twin Earth case differs from Moral Twin Earth precisely because there is no disagreement in the original.[47]

Even if we cannot use the relationship between contents to explain the disagreement-relation in **Primitive Widespread Disagreement**, we can use it in a partial characterization. Some disputes are *verbal*.[48] These involve assertions that have the surface-level grammatical form of sentences that would appear to be inconsistent. For example: a speaker that asserts 'there is money at the bank' appears to assert something inconsistent with someone who says 'there is no money at the bank'. But the appearance of disagreement is not genuine if owing to the ambiguity of 'bank' in English, the second speaker means that there is no money on the side of the river. This is a verbal dispute; the parties to the dispute are talking past one another since they are not using their word 'bank' to make incompatible statements. (Both speakers can agree with the claim made by the other: it is not inconsistent to hold that there is money in the financial institution but no money at the side of the river.) Moreover the dispute is *merely* verbal, because once the speakers realize that the ambiguity in the word 'bank' is responsible for the apparent dispute, they should conclude that there is nothing further at issue between them.

[47]On some views of content, beliefs are implicitly indexed to the world at which they are formed (and, perhaps, some other parameters as well). Such a view would entail that what Putnam's Earthlings and Twin Earthlings believe are contradictory. On such a view Earthlings believe that the stuff in the Pacific Ocean is H_2O *on Earth*. Twin Earthlings believe that the stuff in the Pacific Ocean is not H_2O *on Twin Earth*. It is clear however that this will not serve our purposes as an explanation of why it is that Putnam's Earthlings and Twin Earthlings do not disagree; the exact same line of reasoning would go through for the Moral Twin Earth case from Horgan and Timmons. But in this case, the communities *do* disagree.

Outright *inconsistency* needn't be necessary either: there is a sense in which someone who believes p disagrees with someone who deliberately suspends judgment concerning p. (Friedman, 2013)

[48]Manley (2009)

We can call a dispute that is not verbal a *substantive* disagreement. The surface-level form of a disagreement is not sufficient for a substantive dispute; in addition the parties to the dispute must be disagreeing because they are using their terms to talk about the same thing. That is: unlike in a case where speakers are using 'bank' with different meanings, in a substantive dispute, speakers must use their words with the same meaning; given the simple view of meaning I am working with here, this means that in a substantive dispute, speakers refer to the same entity or property. This is only a necessary condition; shared reference is not *sufficient* for a disagreement. Instead of developing a theory of the sufficient conditions for a substantive theory of disagreement, I will simply note that this is an issue, and rely on an informal, although not entirely natural, grasp of the notion of an inter-communal disagreement.

As a number of authors have pointed out, there are types of disagreement that are not *merely* verbal, but also are not substantive, in the sense outline here. One early example is Stevensonian disagreement in attitude: Stevenson concedes that speakers who disagree in attitude might do so using sentences which are both true at the level of descriptive meaning.[49] Other proposals hold that since normative terms are context-sensitive, speakers can be disagreeing over which context to be in.[50] A related idea holds that the disagreements are meta-linguistic: even if speakers are using the same term with different meanings, they can disagree about which meaning they *should* be using.[51] The notion of *disagreement in plan* in Gibbard (2003) also fits this characterization. Two agents disagree in plan just in case it would not be possible for a single agent to coherently adopt both plans at the same time.[52]

[49]Stevenson (1937, 24), Finlay (2017)
[50]Silk (2017); see the discussion of Kratzer (1977) for the kind of context-sensitivity that makes this negotiation possible.
[51]Plunkett and Sundell (2013), Stroud (2019).
[52]Gibbard (2003, 68 ff.). See also Worsnip (2019) for a generalization of this idea.

I will not explore the possibility of accounting for disagreements across modal space with these non-substantive notions of disagreement. The reason is primarily methodological.[53] The project of the present book is to ask whether a simple realist view can capture the facts about moral and normative disagreement. The most natural interpretation of cases like the Moral Twin Earth thought experiment, for the realist, is an interpretation according to which the relevant communities have substantive disagreements with each other. A realist view which could explain the relevant disagreements as a substantive disagreement will not have to ask for any concessions from opponents who wish to wield Moral Twin Earth-type thought experiments. By contrast, any realist approach which concedes that substantive disagreements are absent in these cases, but tries to make up ground by locating other, non-substantive disagreements between the communities, will face warranted skepticism. Such accounts are strained in conjunction with the realist's metaphysical commitments: if practical language serves to describe what reality is like, then it is highly natural to expect that practical disagreements will be disagreements over what that reality is like. By setting the standard high, and requiring that the disagreements in question come out as substantive, I aim to avoid worries along these lines.

Assuming that moral and normative disagreement is universal, then, the target explanandum is widespread *substantive* disagreement. We can then refine further the thesis that is at issue. It is the **Primitive Ambitious Universal Substantive Disagreement** thesis:

[53]But these non-substantive notions of disagreement do face obstacles. Some are most easily explicated with examples of intra-community disagreement. It is not clear how to extend the characterization of non-substantive disagreement to inter-community disagreements, which needn't involve causal contact or awareness of the communities that are part of the disagreement. Others, which generalize Gibbardian disagreement in plan to other attitudes, risk counting non-disagreements, including the relationship between Putnams Earthlings and Twin Earthlings, as genuine disagreements.

Primitive Ambitious Universal Substantive Disagreement Any two communities that use a primitive term with the same practical role are thereby capable of having a *substantive* disagreement with each other with that term.

As a terminological note, for brevity, I will hereafter speak of **Universal Disagreement**, simpliciter. But this should be understood as shorthand for **Primitive Widespread Substantive Disagreement**, unless otherwise noted. In Chapter 2 I will identify some respects in which disagreement is not *universal*, in the sense outlined here. However, I will retain the qualifications that whatever disagreements between possible users of practical terms does exist should be thought of as primitive, ambitious, and substantive.

1.4 A complication: the Kratzer semantics

There is a complication which we have ignored so far. 'Right', 'ought', and 'should' can in English be used with a practical role. They can also be used roles that are neither moral nor normative. They also have *end-relational* uses ('if you want to reach your destination, you ought to head west on Delmar'), uses that express the requirements of social norms ('you ought to use the outside fork first'), and purely formal uses within a system of rules ('you ought not to move your rook diagonally').[54] Strictly speaking, then, it is ill-formed to ask whether a term in a language has a moral role, or the Gibbard role, etc. These questions only apply to uses of an expression *in a context*.

Context-sensitivity of this kind will need to be accounted for by any theory of practical language. One influential theory, which aims to systematically integrate this phenomenon into a single semantic structure for terms like 'ought' is from Kratzer

[54]See Finlay (2009), Foot (1972), and McPherson (2015), respectively.

(1977), which I will make reference to in various places in this book.[55] (However in cases where the semantic complexity to 'ought' and related terms is not relevant I will ignore the details of the Kratzer semantics.) For Kratzer, all uses of 'ought', 'should' and related terms have a common, context-invariant structure. Each takes a set of options—called a *modal base*—and ranks the options according to some set of standards—an *ordering source*. Which actions are contained in the modal base, and what standards provide the ordering source, are supplied by context. I will refer to this as the *contextualist* account of 'ought'.

On the Kratzer semantics for 'ought', a sentence as used in a context ⌜ought ϕ⌝ is true just in case ϕ-ing is the action in the modal base supplied by the context that ranks highest according to the ordering source supplied by the context.

As an example, take the following sentences:

J John ought not to steal that book from the library;

S Sally ought to hurry in order to catch the next train.

J and **S** can be interpreted in a number of ways, but the most natural interpretations involve different "flavors" of 'ought'. **J** is most naturally read as saying that John *morally* ought not to steal the book from the library—i.e., that morality prohibits John's stealing. **S** by contrast is not making a claim about what is morally required. One might utter **S** when one knows that Sally wants to be at work by a certain time, knows that the next train is about to arrive at Sally's stop, and that Sally will not be at work on time if she misses the next train. Roughly in this scenario **S** says that given Sally's preferences and situation, hurrying the best thing to do.

[55] See Dowell (2011), Chrisman (2015), Silk (2017) for recent applications and developments of this view.

The context in which **J** and **S** are used determines the "modal base", i.e., the options that are available to be ranked. Sally's hurrying to the train is an option for trying to accomplish her goal of getting to work on time. Walking, running, and hailing a cab are also options, but they are not her best options, since they are not as likely to succeed, are more expensive, or more physically taxing, and so on. However there are other courses of action that are not ranked, and not because they would be less effective. Sally's flying to work is not an option; nor is her taking a helicopter. These actions are not in the modal base, in Kratzer's jargon.

Context also determines an "ordering source". The ranking at issue in **S** is prudential: hurrying to the train satisfies Sally's goal of getting to work better than her other options (since she wants to get to work on time). In other contexts *moral* rankings are at issue: in a typical context where **J** is uttered, it is true because morality ranks stealing books from the library below refraining from stealing. Other rankings are possible too: legal norms, etiquette norms, and institutional norms can all supply an ordering source. Two especially salient orderings for our purposes are those that order actions according to standards for what morally ought to be done, and those that order actions according to what all-things-considered ought to be done.

When this contextualist account of 'ought' and cognates is in play, **Universal Disagreement** needs to be refined, but the fundamental explanatory demand on the realist remains unchanged. The relevant disagreements between possible communities are not simply between communities that have a term that has a practical role; the disagreements are to be located in particular contexts of use in each community, where each community is using their term 'ought' with the same moral or normative flavor. I will raise this issue in later chapters, in order to show that the realist view I develop

can accommodated within the contextualist framework. However, to reduce complicating factors, I will only re-raise this issue in specific places; otherwise I will continue to frame the realist view in a simpler semantic framework.

1.5 The central explanatory datum: semantic stability

To conclude this chapter, I will draw out what I take to be the central explanatory task for a realist view, if **Universal Disagreement** is true.

In the simple framework I have set out, the meaning of a practical term such as 'right' or 'ought' (Kratzer-style complications aside) is accounted for by saying what property the term refers to. Two communities use a practical term with the same role are capable of a substantive disagreement just in case they use the same practical term to refer to the same property. A corollary of **Universal Disagreement** in this framework is the thesis that practical terms are *semantically stable:* differences in use between different possible communities do not change what a practical term refers to, so long as practical role is held fixed. This is the **Universal Stability** thesis:

Universal Stability For any possible communities c and $c*$, if c and $c*$ have a tern that is used with the same practical role, then they refer to the same property.

Qualifications analogous to those that we added to **Universal Disagreement** will be relevant here. **Universal Stability** should be interpreted as an ambitious thesis, holding that all possible communities that use a practical term with the same role thereby co-refer, regardless of what other differences there are between them. There are exceptions, since the thesis should cover only *primitive* practical terms, and not those that are understood by their uses to apply to certain kinds of acts by definition.

In the language of Williams (forthcoming): *normative role determines reference*. For

the Ardent Realist in Eklund (2017, 10), stability is a consequence of a *referentially normative* term.

If **Universal Stability** is true, it is a striking thesis. One reason for this is that the moral or normative roles that determine reference, according to **Universal Stability**, are remarkably thin. Take a normative 'ought' that is used by a community with the Gibbard role. This means that the community in question treats an agent as incoherent when they apply 'ought' to an action, and fail to do it. The Gibbard role does not discriminate between communities that routinely apply their normative 'ought' to unfettered pursuit of self-interest, and those that apply the term to selfless acts directed at saving the world. Both communities can consistently use a normative 'ought' in these ways, and use it with the Gibbard role. Stability implies that these differences are irrelevant to what the communities in question are talking about.

From a meta-semantic perspective, it would be surprising if this single aspect of usage was, in all cases, sufficient for determining reference—no matter how the other aspects of usage turn out. A second and related point is that non-practical vocabulary appears not to share to same feature. Some descriptive vocabulary is extremely *un*stable. Take color-terms: if we imagine a community that uses 'red' with the same role as us, we might imagine a community that accepts claims like 'red is darker than pink' and 'red and green are complementary'. But it is extraordinarily easy for such a community to be talking about something other than we do with 'red'—that is, something other than redness. All we need to do is to suppose that this community applies their term red systematically to a different range of reflected wavelengths—for instance that they include some red-orange shades in their application of 'red', and leave out

some of the shades of darker red.[56] In this case it is natural to interpret the other community as talking about a different range of wavelengths—a range that is distinct from redness.[57]

Other non-practical terms are not nearly as stable as **Universal Stability** says practical terms are. Horgan and Timmons gesture at one source of the absence of stability for descriptive terms with the explicit reference to Putnam's original Twin Earth case. As Putnam's case shows, 'water', even when used in roughly the way English speakers use the term, refers to different substance in different environments. In an environment where there is no H_2O, but where there is XYZ that behaves much like water at a macroscopic level, speakers use 'water' to refer to XYZ. This is a kind of failure of stability, since speakers using 'water' in similar ways fail to refer to the same thing.[58]

With many ordinary descriptive terms, it is easy to have the feeling that disputes involving the term are merely verbal. One example comes from Manley (2009). Take a pair of speakers who have a dispute about what constitutes a cup by asserting the following sentences when referring to the same piece of glassware:

A : This glass is a cup.

B : No, it isn't—cups aren't made of glass.[59]

It is possible that, given the meaning of 'cup' in English, one of these speakers is correct. But this applies only to an intra-community dispute about cups, namely one

[56]For concreteness, we can imagine that whereas we apply 'red' to reflected wavelengths between 620–750 nm (with some hesitation for borderline cases), the alternative community applies 'red' to reflected wavelengths between 600 – 730 nm (with the same hesitation for borderline cases). Of course we will also need to imagine compensating shifts in use of other color-terms to produce a compelling example of the plasticity of color-terms.
[57]Cf. the "permutation problem" in Smith (1994).
[58]Dunaway and McPherson (2016) explores this in greater depth; the phenomenon is due to the fact that the candidate referents for 'water'—viz., H_2O and XYZ—are not instantiated in all the same worlds. This is a failure of "intensional similarity".
[59]Manley (2009, 10)

in which both of the participants are English-speakers (or speakers of some language very much like English). If A and B are from different communities, where other members of each community share their dispositions to use 'cup' (and otherwise speaks a language that is exactly like English), then it is not hard at all to interpret A and B as speaking about different things. Perhaps A is talking about cups, and B is talking about a distinct kind of thing, namely cups-not-made-of-glass. An inter-community dispute like this would be merely verbal. 'Cup' is not very stable.

Perhaps the apparent stability of practical terms does reveal some deep incoherence in the realist view. It is an open possibility that, while a fully adequate account of practical terms should entail **Universal Stability**, no account that appeals to realist resources—that is, no account which treats practical terms as referring to part of reality—can explain it. If so, we would be faced with the option of either adopting a truncated form of realism which does not have ambitions to explain the relevant stability-facts,[60] or rejecting realism in favor of non-cognitivism or error theory.[61]

But before investigating what a realist can explain, we should ask more carefully what a realist *should* explain. In Chapter 2 I argue that **Universal Disagreement** and **Universal Stability** are not true. Horgan and Timmons and others point in various ways to a remarkable degree of stability for practical terms. But we should not accept, on the basis of their examples, that the extent of the stability is as broad as **Universal Stability** claims. By raising cases which falsify it, we can set an important limit to the explanatory target for any meta-ethical view, including that of the realist. It should explain the extent of the stability for practical terms, but should also explain why there

[60] See for example Railton (1986) and Brink (2001).
[61] The "what's at issue" argument in Gibbard (2003) appeals to facts in the vicinity, though these are intra-community disagreement facts. Error theorists such as Streumer (2017) deny that practical terms refer to anything, so will be forced to be revisionary about a stability thesis.

are limits short of what **Universal Stability** claims.

Chapter 2.

Failures of Stability

Chapter 1 raised the issue of disagreements involving practical terms. Disagreements show up in some familiar cases, such as Horgan and Timmons's Moral Twin Earth case, and can be extended in a number of ways. One question that arises is: how far does the disagreement extend? The spatial metaphor can be fleshed out using the relationships between different possible linguistic communities. If possible linguistic communities such as those found in the Moral Twin Earth case disagree, which other possible communities disagree with each other as well? If, as the realist should think, these disagreements are *substantive* disagreements, the question concerns the stability of practical terms: which possible communities use their moral terms to refer to moral rightness, and which use a normative term to refer to what is all-things-considered obligatory?

One hypothesis is that two communities refer to the same property if they use a term with the same practical role. For example: any two communities that have a term 'wrong', and use it with a moral role by connecting feelings of blame and guilt to its application, will be referring to moral wrongness. This is a natural hypothesis given the facts of the original Moral Twin Earth case and nearby variants: the linguistic communities in question both use a term with the same role, but differ in a number of other respects, including which actions they apply their moral terms to. This is the **Universal Stability** thesis, which holds that all possible linguistic communities that use a term with the same practical role will refer to the same property. If it is true, then it supports the **Universal Disagreement** thesis.

Universal Disagreement characterizes one potential explanatory task for realist theorizing. In this chapter I set out what kind of explanation is in order: this is a a job for a *meta-semantic* theory; the realist needs a general theory of what determines the reference for terms in a language that explains the semantic facts. §1 elaborates on this general task. The rest of this chapter is dedicated to showing that **Universal Stability** is not, despite its simplicity and naturalness, the target explanatory goal. §2 outlines the potential for a weaker thesis which still explains the specific examples of disagreements with practical terms raised by the Moral Twin Earth case and variants. I call the weaker hypothesis **Robust Stability**. §3 argues that there are a number of structural possibilities for the failure of **Universal Stability**, and cases modeled on the original Moral Twin Earth example do not rule out counterexamples that fit these structural patterns. §4 argues that these are more than mere in-theory possibilities: there are specific examples that, following the same methodology of the Moral Twin Earth case, show that **Universal Stability** is false. SS5-6 extend these lessons to the contextualist Kratzer semantics and draw some more general conclusions.

2.1 Explanations of disagreement

The thesis we are calling **Universal Disagreement** is shorthand for the **Primitive Ambitious Universal Substantive Disagreement** thesis:

Primitive Ambitious Universal Substantive Disagreement Any two communities that use a primitive term with the same practical role are thereby capable of having a *substantive* disagreement with each other with that term.

I have primarily used **Universal Disagreement** as an expository tool: much of the surrounding literature emphasizes the extent of disagreements involving practical

terms as used by different communities across modal space. Any limits to the range of possible disagreements are not emphasized; the focus is on the extent to which such disagreements do occur.

If **Universal Disagreement** is true, then a realist will want to explain this by providing a meta-semantic theory which entails **Universal Stability**. There are a few examples in the literature where **Universal Stability** is explicitly taken on board.

Wedgwood (2001) makes this point by imagining two agents ("friends") who have the same view of the facts about a situation, but differ in their fundamental moral intuitions. The difference in intuitions does not give rise to a difference in the role with which they use their moral terms—Wedgwood is gives a specific account of the role of moral terms which makes it clear that speakers who accept very different specific judgments about which things are right and wrong can do so while using their moral terms with the same role.[62] Holding fixed this role, he says *any* difference in intuitions will give rise to disagreements:

> According to my semantics, you and your friend both mean the same thing by the term 'wrong', because you both master the rule according to which certain moral beliefs commit one to a certain sort of preference or endorsement of attitude. That is, for both of you, sincere acceptance of a sentence involving 'wrong' has the same consequences for practical reasoning—even if your moral thought is guided by different fundamental moral intuitions, so that you form opposite moral beliefs on the basis of the same nonmoral beliefs. (Wedgwood, 2001, 29)

The disagreement that Wedgwood explains will be substantive: if, as he claims, any two speakers who use their term 'wrong' with the same role mean the same thing, then on his semantics they refer to the same property. Wedgwood's explanation is not

[62] In Wedgwood (2001), the fundamental moral terms is 'better', which has the following role: anyone who accepts $\ulcorner A$ is better than $B\urcorner$ is committed to preferring A to B (p.15). Wedgwood (2007, Ch. 4) gives a related semantics for 'ought'.

limited to one specific example: *any* instance of shared role with different intuitions will produce a substantive disagreement.

Williams (forthcoming) explicitly proposes a theory that entails **Universal Disagreement** for the moral 'wrong'. The central piece of machinery in Williams's theory is the claim that a term refers to the property that makes use with its distinctive role most *reason-responsive*.[63] A use of a term is like any other act—it can be supported by reasons, or there can be no (or bad) reasons to do it. According to Wedgwood's meta-semantics, 'wrong', when used with the moral wrongness role, refers to whatever property would give us reason to, in part, blame agents who perform acts that 'wrong' applies to.

There is, we can assume, some property that gives us reason to blame agents who perform acts which have this property. This is a claim that a substantive theory of normativity will make: it isn't uncontroversial, but barring significant normative skepticism, we will concede that there is some property possession of which gives us reason to blame agents for performing acts that have it.

Adding a substantive normative assumption about a genuine reasons to use a term with the moral wrongness role produces an explanation of **Universal Disagreement**, applied to moral terms. Suppose there is some property P that gives us reason to use 'wrong' with the moral wrongness role—that is, to blame agents who perform acts instantiating P, and refraining from blaming agents who perform acts which lack P. A particular substantive theory of broad normativity will entail that a community that uses 'wrong' with the same role will not have reason to use 'wrong' differently. That

[63] In some ways this builds off of Wedgwood's claim that a term refers to the property that makes use of the term with its "essential conceptual role" *correct*. (See e.g. Wedgwood (2007, Ch. 4).) Since Williams does not make as many assumptions about the metaphysics of the role in question (by not, for example, assuming that each term is associated with a concept for which some aspects of its role are "essential") I will focus in more detail on Williams's machinery.

is, if we have reasons to blame agents who perform acts with property P, then likewise other agents have the same reasons to blame agents who perform P-acts, even if they use their term 'wrong' differently.[64] It follows that both communities who use 'wrong' with the moral obligation role are both referring to P.

The explanation (Williams claims) generalizes, and so **Universal Disagreement** for 'wrong' follows. 'Wrong' when used as a moral term refers to whatever property rationalizes its use with the moral wrongness role—that is, with its characteristic patterning with application and withholding of blame. Take the Moral Twin Earth scenario as an illustration: for a community that uses 'wrong' as a Utilitarian would, there is a fact about whether their application of blame to actions that are not happiness-maximizing are reason-responsive or not. We need not assume that they are: this is a question for substantive normative theory. But suppose that these applications are reason-responsive, and they have good reasons to blame any agent that does not perform a happiness-maximizing act. Then a community of deontologists, who apply 'wrong' to acts that constitute violations of autonomy, *also* have the same reasons. They always have reason to blame agents who do not perform happiness-maximizing acts. They sometimes fail to do what they have reason to do, but the applications of 'wrong' that would be reason-responsive are the same for the two communities. And so they are both talking about the same property, failing to maximize happiness.

The same story could be repeated for any community that uses 'wrong' with the moral wrongness role. Williams says that these cases support the conclusion that 'morally wrong' displays universal stability, rather than a more limited thesis (here, he is using 'referential stability' to denote a thesis like **Universal Stability**; 'W' stands for

[64] Again, this is a consequence of a substantive theory of normativity, and is not trivial. Williams supplies additional details to secure this result in Williams (forthcoming, 104-5). The theory, given alternative substantive theories, will produce different verdicts.

moral wrongness, and the 'blame-centric role' is what I am calling the moral wrongness role):

> A specific "twin earth" thought experiment such as this supports an instance of the universally quantified referential stability thesis, and referential stability in full generality will be supported if analogous verdicts are accepted for other communities (including ourselves) who have a concept W that plays the blame-centric role. (Williams, forthcoming, §4.2)

The "universally quantified referential stability thesis" is, like **Universal Stability**, a thesis about *all* possible linguistic communities who use a term with the same moral role. I don't intend to endorse Williams's substantive meta-semantic theory (I provide an alternative in Chapters 3 and 4) or his argument that his theory entails something like **Universal Stability**. Rather the point here is simply that **Universal Stability** appears in Williams, as in Wedgwood, as a plausible explanatory desideratum. They provide meta-semantic theories designed to explain it, and by extension, **Universal Disagreement**. It is worth asking whether this target is what a meta-semantic theory should be aiming for.

2.2 Revisiting Moral Twin Earth

In Chapter 1 we sketched the Moral Twin Earth case and some variants. Although the original case, as presented by Horgan and Timmons, is (as Williams says) just one instance of the general **Universal Stability** thesis, it is also a guide to producing more instances. We can produce additional instances by varying certain features of the original case.

As we noted in Chapter 1, additional instances can be generated is by imagining communities that have different substantive theories of morality or normativity. Horgan and Timmons provide an example which involves one community that accepts a

consequentialist theory of morality, and a second community which accepts a deontologist theory, and do not need to specify the details of the respective theories. In part this is not necessary: it is intuitively clear that there is an instance of **Universal Disagreement** involving communities that accept any of the specific versions of the consequentialist and deontological moral theories. In addition we do not need to limit ourselves to substantive differences that have analogues in the consequentialism-deontology debate. A community resembling virtue theorists, who use their moral terms 'right' and 'wrong' to track what an agent with certain virtues would (not) do, will have substantive disagreements with other possible users of moral language.

This is a recipe for extending the Moral Twin Earth cases while preserving intuitions of disagreement: we simply need to consider communities who use their terms with the same moral role, but differ over which actions they apply 'right' and 'wrong' to owing to adherence to different substantive moral theories. It is very plausible that these communities will appear to disagree with each other. Since there is no incoherence in a description of a community that blames and feels guilt for the actions proscribed by almost any conceivable moral theory, such communities will count as one of the communities in modal space that disagrees with others who possess a moral term.

This does not amount to a vindication of **Universal Disagreement**. The reason is straightforward: **Universal Disagreement** makes a claim about *all* communities who use a primitive moral term with the same practical role, but differ over other aspects of use with such terms. But the communities that differ only over substantive moral theory do not exhaust the range of communities covered by **Universal Disagreement**. A full evaluation of the thesis should cover the leftover cases.

Some terminology will help here. One thesis is that, among possible communities that use a practical term with the same moral or normative role, all those communities that differ *only* in which substantive moral or normative theory they apply their terms with will be capable of having a substantive disagreement. Call this the **Robust Disagreement** thesis:

Robust Disagreement Any two communities that use a primitive term with the same moral or normative role, *and differ at most in which substantive theory they follow in applying the relevant term*, are thereby capable of having a substantive disagreement with each other with that term.

Robust Disagreement is weaker than **Universal Disagreement**, since it does not make any claims about possible communities who differ in more than the substantive theory they accept. **Robust Disagreement** is also a thesis that is supported by the variations on the original Moral Twin Earth thought experiment outlined above. Accepting the usual intuitions about the Moral Twin Earth case, then, does not directly force us to accept **Universal Disagreement**. A meta-semantic theory that explains **Robust Disagreement** but not **Universal Disagreement** would be adequate to explain the intuitive force of the examples that have been raised.

Even if we set the goal of explaining **Robust Disagreement** and not **Universal Disagreement**, the explanatory task would still be significant. **Robust Disagreement** requires that practical terms are highly stable, and plausibly requires that they are much more stable than most descriptive vocabulary. So even if we were to accept a retreat to **Robust Disagreement**, we would not have avoided all significant explanatory burdens.

Why accept **Universal Disagreement** and not just **Robust Disagreement**? This

would be motivated by considering examples of possible communities who differ not just in the substantive theory they follow when applying their moral or normative terms. It might be that, even with these additional differences in usage of practical terms in place, it is still compelling that the communities in these cases have substantive disagreements. But the existing range of examples that can be obtained from simple variations on the Moral Twin Earth case do nothing to support this conjecture.

I will argue in what follows that this conjecture is false. By focusing first on a structural characterization of the kinds cases that would support **Universal Disagreement** and not **Robust Disagreement**, we can fairly easily locate counterexamples to the stronger thesis. The same kind of intuitive reflection on cases that supports the judgment that there is disagreement in the original Moral Twin Earth case strongly suggests that there is *no* disagreement in cases that would be needed to support **Universal Disagreement** and not just **Robust Disagreement**. This will force us to substantially revise the explanatory target for a realist meta-semantics. It will need to account for a measure of semantic stability that is induced by a practical role. But it cannot do this by holding that the practical role is the only semantically relevant use of a practical term.

I will argue as follows. In §3 I present an outline of the structural features of several potential counterexamples to **Universal Disagreement**. These structural outlines do not by themselves count as an argument against the thesis. Genuine counterexamples will need to take the form of relatively specific characterizations of possible linguistic communities that meet the following conditions: they use terms which share a practical role, but fail to talk about the same thing and be capable of disagreeing with each other. In §4 I turn to the concrete counterexamples, which are constructed on the outline of

the structural features sketched in §3. The closing sections of this chapter clean up some loose ends and complications.

2.3 Shared role without disagreement: structural features

2.3.1 Role strengthening

Let *R* designate a specific practical role. This could be the Gibbard role, or some moral role. **Universal Disagreement** says that every community which uses 'ought'[65] with the role *R* is thereby talking about the same thing, namely obligation, and is capable of disagreeing with any other community that uses 'ought' with role *R*. But use with the role *R* is compatible with lots of variations in use of 'ought' along other dimensions.

Robust Disagreement is a limited version of **Universal Disagreement**; according to the more limited thesis, any two possible communities that use 'ought' with *R* will disagree with each other so long as the only differences between them consist in which substantive moral or normative theory they follow when applying their 'ought'. The contrast between this thesis and **Universal Disagreement** is instructive here, because while the original Moral Twin Earth case and nearby variations all support **Robust Disagreement**, they do not on their own support the further idea that there is disagreement in the additional possible communities that **Universal Disagreement** ranges over. But such possible communities do exist, because there are ways for two communities to both use 'ought' with role *R*, yet to differ in other uses of 'ought' which do not amount to simple differences in which substantive moral or normative theory their

[65]For the sake of simplicity I will continue to write as if each community that uses practical language uses the strings 'ought', 'right', etc. with the relevant practical role. This is a simplification, and the general point is not limited to communities that use a term that is written with any specific string of inscriptions.

use conforms to.

Some of these uses are found in a possible community which uses their 'ought' with *additional* roles. For example: suppose one community (we can suppose it is the actual community of English-speakers) uses their 'ought' with the role R, and thereby refers to the property P. Another possible community uses their 'ought' with role R, and also use 'ought' with an additional role. The additional role is one that characterizes the property P^+, where the objects that instantiate P^+ are a subset of the objects that instantiate P. There is no inconsistency in using a practical term with both roles: finding a referent that satisfies R does not ipso facto involve a referent that does not satisfy the additional role. The additional role can make additional semantically relevant contributions to what 'ought' in the mouths of the possible community refers to; it is in principle possible that such a community refers to P^+ and not P, and so refers to a different property than we do. We can say that a role role with this semantic effect *strengthens* R.

An analogy with logical connectives applied to predicates may be helpful here. Take a predicate F which refers to the property F-ness. Using logical connectives to combine F with other predicates produces an expression with a different referent: $\ulcorner F \wedge G \urcorner$ refers to a property that has the objects at the intersection of F-ness and G-ness as its extension. This is a property distinct from F-ness, assuming F-ness and G-ness are distinct. The conjunction strengthens F in an analogous manner to the strengthening of R by the additional role.

Thus it is possible that a good meta-semantic theory should entail that a practical term, when used with an additional role, refers to P^+, the property that best fits the role that does the strengthening. P^+ has an extension that is a subset of P. This is

not on its own an argument that the are possible uses of practical terms that involve additional strengthening roles. The point here is just that the concrete cases which support **Robust Disagreement** do not rule out the possibility of strengthening roles, which have a semantic function analogous to conjunction introduction for predicates.

2.3.2 *Weakening*

If strengthening roles are possible then the converse phenomenon should be possible as well. Take a community which, given the totality of facts about their use of a practical term 'ought', including use with the practical role R, refers to the property P. A second possible community might have a practical term with the same practical role R, and yet not include other aspects of the first community's role-like uses in their own. There is then another property P^-, distinct from P, which includes some acts that are not in the extension of P. If the second community uses 'ought' with a role best fit by P^-, then they use their their practical term with a *weakening* role.

2.3.3 *Overriding roles*

Conjunction provides an imperfect analogy for the semantic function of strengthening roles. The point of disanalogy lies in whether the semantic features can be read off of the structure alone (as in the case of conjunction), or whether it is merely possible that the additional structure has a semantic effect. In order to show that a meta-semantic theory should assign a different referent to some normative terms used with additional roles, we need to produce some examples where this seems like the correct verdict. I provide some in the next section. But the disanalogy between meta-semantics and logical constructions points to another kind of counterexample to **Universal Disagreement**. While use of a term with a specific role contributes to some

degree in determining what the term refers to, any particular use is not decisive: a particular aspect of usage can be *overridden*. (Some English speakers apply their term 'fish' to sharks, but this doesn't imply that on some occasions the English word 'fish' applies to sharks.) This is why we cannot count on the mere structure of strengthening roles to produce counterexamples; it is always possible that normative or moral roles override these additional roles.

If **Robust Disagreement** is true, these roles already do a lot of overriding—so long as a possible community has a practical term, no amount of difference over substantive theory will produce a difference in reference.[66] The role overrides any amount of fit with usage in accordance with a particular substantive theory. But the possibility of overriding cuts both ways. It is possible that some terms are used with a practical role, but this usage is overridden by other role-like aspects of their use.

More concretely: suppose that P is the property is property that fits our use of a practical term 'ought' with R. Even so, there might be possible communities who use 'ought' with R, but also use 'ought' with an additional role that overrides R. This means that 'ought' when used in this way refers to the property $P^{\dagger}$, where the extension of $P^{\dagger}$ overlaps with, but is not identical to, the extension of P.

Just as with potential strengthening roles, we cannot read off of this structural description alone which communities are referring to something other than what typical users of practical language are referring to. Rather we need particular examples involving these additional roles, which give rise to intuitive judgments that practical terms in these examples have different referents. But, even with the cases that motivate **Robust Disagreement** in hand, we cannot assume that the meta-semantic facts will go

[66] If **Universal Disagreement** is true, then moral and normative roles are *always* overriding in this sense.

similarly for communities that use their practical terms with potentially overriding roles.

Here is a (somewhat contentious) example of an overriding role in a non-practical case. We can imagine a community of speakers who speak a language that is similar to English, and who use most of their terms like we do. One aspect of their use of the word 'mammal' is characterized by a role they articulate with the sentence 'mammals give birth to live young'. There are various options for how we should characterize the exact status of the mammal-role. But this is not the only aspect of usage that determines what 'mammal' picks out: a platypus is a mammal, but lays eggs instead of giving birth to live young. Nevertheless the connection to birth of live young is an important part of the mammal-role in their language, even if it does not exhaust the reference-determining features. Plausibly other roles for 'mammal' explain this: mammals also have common genetic and evolutionary traits, which the platypus shares with other mammals.[67] If it is clear that aspects of use of 'mammal' in this community make the evolutionary and genetic features are relevant to the reference of the term 'mammal' in their language, then these speakers are an example where the role connected to birthing live young is overridden.

Practical terms are, potentially, no different. Even if they are used with a distinctive moral or normative role, there is the possibility that some community will use the terms with additional roles that partially override the normative role. Use of 'ought' with a practical role does not exclude the possibility of uses with additional roles. Some of these might bear on the facts relevant to reference-determination for practical terms. The next section provides some examples to suggest that this is in fact the case.

[67] https://en.wikipedia.org/wiki/Platypus

2.4 Shared role without disagreement: examples

There are examples of each type of structural feature outlined in the previous section: strengthening roles, weakening roles, and overriding roles. I will focus on providing an example of each which using 'ought' as a normative term. I will note briefly how similar examples arise for terms used with a moral role as well.

2.4.1 Weakening

Begin with a weakening role for a normative 'ought'. Assume for simplicity that as English speakers we use the normative 'ought' with the Gibbard role. This means that we treat it as playing a deciding role in deliberation, holding that it is a conceptual mistake to judge, using the normative 'ought', that one ought to ϕ, and then fail to ϕ.

This is not the only role-like usage associated with the normative 'ought' in English. Our actual use of the normative 'ought' treats agents who are not able to perform an act as not obligated, in the normative sense, to do it. That is: if agent a cannot ϕ, then we do not apply 'ought' to a's ϕ-ing. Call this the *ability role*.

There are many questions about the ability role, since it is related to the doctrine that 'ought' implies 'can', and this question has inspired a large philosophical literature. I will not dive into the details because they are not important for present purposes. But a few notes on why this is so are in order. First, much of the debate is not concerned with sociological facts concerning how English-speakers in fact use the term 'ought'. Instead, the question is a normative one, concerned with how a theoretically interesting normative notion is related to ability. Second, the answers to this normative question are not obvious (otherwise a large philosophical literature would not exist) and are not necessarily reflected in actual practice by English speakers. The

claim that in English the normative 'ought' is used with the ability role does not entail that English speakers have settled on a specific claim about the precise relationship between normative obligation and ability. Rather, we can understand the claim that 'ought' in English has the ability role as a claim that there is some notion of ability that places semantic constraints on what the normative 'ought' applies to.[68]

We can think of these English-speakers who use their normative 'ought' with both the Gibbard role and the ability role as analogous to the speakers on Earth in the original Horgan and Timmons thought experiment. This specification leaves open questions of substantive normative theory; for instance we can either add that in addition these speakers all agree on some consequentialist normative theory, or allow that in this case there is no unanimous conformity to a particular substantive normative theory.

Now consider an analogue to Moral Twin Earth, which we can call *Ability Twin Earth:* a community of speakers who use their term 'ought' with the Gibbard role, but do not use it with the ability role. Since this community does not use their term with the ability role, they do not restrict their applications of 'ought' to actions that an agent has the ability to perform. Instead, they apply the term to whichever actions would be best for an agent to perform, regardless of her metaphysically contingent limitations. We (on Earth) would say things like 'it is not the case that Sally ought to provide (on her own) famine relief for a 500,000 people', because Sally does not have the ability to provide famine relief to this extent. The Ability Twins assert instead 'Sally ought to provide (on her own) famine relief for a 500,000 people' because it is only a contingent limitation on Sally that she cannot accomplish this, and it would be best if she were to

[68]This can be understood in terms of the conditions for role-hood outlined in Chapter 1. Any reasonably specific thesis about the relationship between obligation and ability will not be encoded by unanimous community-wide, psychologically robust usage. But some connection to ability will be.

provide famine relief on such a large scale.[69]

Since both communities are using their term 'ought' as a normative term, this means that both communities use the term with the Gibbard role. That is, both communities treat an agent as incoherent when that agent accepts that the normative 'ought' applies to an action in their circumstance, and yet fail to perform the action. Since the communities differ in whether they use the normative 'ought' with the ability role, there will be further differences between them. In particular, Ability Twin speakers will accept that there are actions the normative 'ought' applies to in their own circumstance, but which they have no ability to perform. Since these speakers will typically fail to perform the acts they have no ability to perform, they will be treated as incoherent by other speakers, as if they are making a conceptual mistake. The mistake is not in applying the normative 'ought' to the act of eliminating the famine; rather the mistake is applying 'ought' to this act *and* not performing it.

This is a perhaps odd feature of the Ability Twins. The community-wide judgment of incoherence in these cases is (in a sense) unavoidable, since once one of the Ability Twins applies 'ought' to ending the famine, owing to the goodness of ending famines, that Twin cannot avoid incoherence, since there is no way for the Twin to actually end the famine. Instances of unavoidable self-ascribed incoherence do not make them impossible. We are familiar with actual examples of people with incoherent sets of credences, and those who do not accept the logical consequences of their own beliefs.

[69]We can imagine the case in more specificity by making assumptions about what theory the Ability Twin community accepts about the action that would be best. They could, for example, accept a theory that resembles a consequentialist theory—providing the famine relief would be best, on this theory, because it is better than not providing relief to the same levels. (Moreover it is impossible to provide famine relief to more than a 500,000 people: according to the Food and Agriculture Organization of the United Nations, around 815 million people globally suffer from chronic hunger (`http://www.fao.org/state-of-food-security-nutrition/en/`). Alternatively we could imagine that the Ability Twins accept some side-constraints on maximizing, but providing famine relief for a 500,000 people would not require violating these side-constraints.

Incoherence is fairly common in the actual world; the Ability Twins in this case simply think they run across practical incoherence on a regular basis.

There is a crucial difference between the Ability Twins and the Twins in simple variants of the original Moral Twin Earth case. The Ability Twins are most naturally interpreted as talking about the best state for the world to be in; or, more precisely, they are talking about what would be best regardless of the contingent limitations of agents. I will call this the *best state property*. When they say 'Sally ought to provide (on her own) famine relief for a 500,000 people', they say this on the basis of the fact that the world would be better if there were no famine, and so they mean that the world would be better if famine relief were provided for 500,000 people. They accept similar claims for other agents: although there are few if any agents who could succeed in successfully providing the relief on their own, this does not matter to the Ability Twins. All that matters for whether they accept such a sentence is whether it would be better for the world to be in the relevant state, *if* it were possible for an agent to bring it about.

The Ability Twins do not appear to have a substantive disagreement with speakers on Earth who say of Sally in the exact same circumstance 'it is not the case that Sally ought to provide (on her own) famine relief for a 500,000 people'. Speakers on Earth do not deny that the world would be better if famine relief were provided for a 500,000 people.They agree with what the Ability Twins say using 'ought' in their own language. But since speakers on Earth use their normative 'ought' with the ability role, they are talking something different. They are talking about what Sally is obligated to do, in the ordinary sense, when they say 'it is not the case that Sally ought to provide (on her own) famine relief for a 500,000 people', and would correctly add 'because she

does not have the ability to provide famine relief to so many people on her own' as a reason for concluding that Sally does not have this obligation.

The speakers on Earth and Ability Twin Earth have a merely verbal disagreement. The claims they make with the expression 'ought' have the form of inconsistent claims: speakers on Earth assert a string with a negation in front of a string speakers on Ability Twin Earth assert. But since 'ought' in the Ability Twins' mouths means something different, they are not actually disagreeing. The Ability Twins are making a claim about the best state property; we make a claim about normative obligation. Since each community uses 'ought' with a normative role, yet fail to substantively disagree, this is a counterexample to **Universal Disagreement**.

An alternative version of Ability Twin Earth with moral terms follows the same pattern. We can imagine that the Ability Twins have a primitive moral term 'right' that they use with the moral rightness role: they blame agents who fail to perform actions that they apply 'right' to, and feel guilty when they themselves fail to perform such actions. By not using 'right' with the ability role, they routinely blame others for not performing actions that they have no ability to perform, and likewise feel guilty for performing such actions. There is nothing incoherent about this, although guilt and blame are much more common in Ability Twin Earth. But given an appropriate description of what kinds of actions the Ability Twins apply their 'right' to, it would appear that they are using their moral vocabulary to talk about what would be the morally best way for the world to be. These are claims that speakers on Earth, who use their moral term 'right' with the ability role, do not disagree with.

The Ability Twins use their practical terms—either moral or normative—with a weakened role. This means that there is a role we use our practical terms with on

Earth, namely the the ability role, that applies to certain kinds of action, namely the actions we are able to perform. On Ability Twin Earth, speakers use their practical terms with a role that is satisfied by a broader range of actions: any action that could be performed by an agent with none of the contingent limitations of humans satisfies this role. This represents a weakening of usage on Earth. In this case the weakened role appears to make a difference to what speakers who use their practical terms with these roles are talking about. There is no guarantee that other examples will produce the same semantic effect. But there is at least one example where it does—and this is enough for a counterexample for **Universal Disagreement**.

2.4.2 Strengthening

A strengthening role for the normative 'ought' involves use of the term with an additional role that is satisfied by fewer actions than the actual roles our normative terms are associated with.

We can say that a community's use of a term has the the *psychological feasibility role* if the community in question applies the term only to actions that a typical agent can perform without experiencing psychological distress or other kinds of discomfort that would make it somewhat unlikely that the agent performs the act in question. Thus, in a case where giving a significant amount of money to charity would cause a typical agent a significant amount of psychological pain, the act of giving to charity does not satisfy the psychological feasibility role. Psychological feasibility plays a part in psychological explanations of behavior: when explaining why an agent failed to give money to charity, the fact that giving the relevant sum of money was not psychologically feasible can (partially) explain why she did not give the money. The

acts that are psychologically feasible are a subset of the acts that we have the ability, in some normatively relevant sense, to perform.

A normative 'ought' can be used with the psychological feasibility role. A typical consequentialist in the actual world will apply her term 'ought' to the action which, among the available options, produces the best consequences. For a person of normal means, giving $5 to a particular charity is among her options. Giving $100, $1,000, $5,000, and $10,000 are also options. In each case, the more money that is donated to charity, the better: every dollar donated (up to an amount much greater than $5,000) does more good when donated to charity than when used at the discretion of a person with normal means. But not all of these options are psychologically feasible. It is plausible that for someone of normal means, giving $10,000 is not feasible in this sense. While it is possible for a person of such means to live on $10,000 less than her normal income, and so in some sense an option for her, it is not something the typical person can do without experiencing significant psychological distress. This fact would be part of a psychological explanation for why, in general, most people do not give $10,000 to charity, even when they are aware of the consequences and could survive on the reduced income.

There is a possible community of Twin Earthlings who use their normative 'ought' with the psychological feasibility role, and thereby use their normative term differently than a community of English-speakers on Earth that use their normative terms without the role. These are the *Feasibility Twins*. On Earth we allow that sometimes we are required to do things that are difficult for us to do, and so doing what is required involves experiencing significant distress. The Feasibility Twins do not speak in the same way: they never view their normative term 'ought' as applying to acts that cause

the relevant levels of distress. Nonetheless we can imagine that the Feasibility Twins agree at some level on substantive questions: each community always applies their term 'ought' to the action which, out of the relevant options, produces the most good. The difference between these Earthlings and Feasibility Twins lies in their views about which options are relevant: for the Feasibility Twins in this case, who use 'ought' with the psychological feasibility role, it is only the psychologically feasible actions that are ranked.

While applying their 'ought' with the psychological feasibility role, the Feasibility Twins also use it with a normative role. Taking the Gibbard role as our characterization of the distinctive role of normative terms, this means that the Feasibility Twins treat those that apply their term 'ought' to their own performance of an action and yet fail to perform it as incoherent. Since they differ from the Earthlings in their use of 'ought' with the psychological feasibility role, they apply the normative 'ought' to different actions in some cases. When there is an action that would produce the most good, but is not psychologically feasible—for instance the action of giving $10,000 to charity—speakers on Earth say 'one ought to give the $10,000 to charity' while speakers on Twin Earth say 'it is not the case that one ought to give the $10,000 to charity'.

It is important to emphasize that while the Twin Earthlings in this case apply their normative term 'ought' to actions that are both obligatory and psychologically feasible, they do not conceptualize application of their normative term in this way. That is, they do not have a term for obligation, and think of the referent of 'ought' in their language in conjunctive terms, applying to actions that are obligatory *and* psychologically feasible. Instead, their normative 'ought' is primitive. There are systematic and robust aspects to the usage of 'ought' by the Twins that qualify as role. But these uses

are not constrained by the uses of distinct terms which form a definition of 'ought' in Twin English. They have no normative term that we would translate as meaning what 'ought' means in English. We should think of the Feasibility Twins' use of an 'ought' with both the Gibbard role and the psychological feasibility role as our theoretical characterization of the pattern of application of their usage; it is not a description of the psychological reality of users of normative language on Twin Earth; nor is it a theoretical description they would (or could) give of themselves.

Moral terms can be used with the psychological feasibility role as well. The term 'wrong' is used as a moral term by a community when it has the moral wrongness role: that is, if speakers systematically blame those who fail to perform actions that they apply 'wrong' to, and feel a corresponding moralized guilt when they themselves fail to perform such actions. As before, we can imagine a version of the Feasibility Twins who use a term with this role, and in addition only apply 'wrong' to actions that are also psychologically feasible. We might then imagine a Feasibility Twin saying, of an act of embezzlement which is extremely tempting because of the benefits for the embezzler's family, 'embezzling the money is not wrong'. Ex hypothesi the Feasibility Twin uses 'wrong' with the psychological feasibility role, and the facts about the benefits of embezzling might be so great in this case that refraining would cause significant distress in normal people. On Earth, speakers say of similar acts 'embezzling the money is wrong', since their use of the term is not accompanied by the psychological feasibility role.[70]

[70]There is an additional complication if we characterize the role for moral terms moral terms in the manner of Williams (forthcoming). There, Williams includes in his characterization of a wrongness-role room for *excuses:* agents who do something wrong do not deserve blame if they have an excuse for performing a wrong action. If we extend the moral wrongness role in this way, we should not view the Feasibility Twins as speaking about moral wrongness, and simply deploying a specific view about the relationship between psychological feasibility and excuses. A community that views psychological infeasibility as an excuse does apply 'wrong' to some acts that are not psychologically feasible. They simply think that these are cases where the wrongness of the act in question exculpates, and so blame is not appropriate. They might

Return to the normative case. The **Universal Disagreement** thesis entails that both we and the Feasibility Twins use a noramtive term 'ought' to refer to the same property. We would have a substantive disagreement about whether giving $10,000 to charity in certain circumstances instantiates the property of obligation, according to **Universal Disagreement**.

This does not seem to be the right result. As users of a normative 'ought' on Earth, learning that Feasibility Twins systematically limit their use of 'ought' by not applying it to psychologically infeasible actions should not make us think that they have a theory of obligation, according to which only psychologically feasible acts can be obligatory. Instead, we should think that are talking about something different than what we are talking about—call this the property of *psychologically feasible obligation*. We, on the other hand, are not talking about psychologically feasible obligation—our term 'ought' refers to obligation, full-stop. If we imagine, as a heuristic, an in-person dispute with the Twins, we might attempt to convince them to use a different term, which does not refer to psychologically feasible obligation. But we would not say that they are speaking falsely in their own language when they make assertions, using their own 'ought', that do not apply the term to psychologically infeasible acts.

These cases constitute additional counterexamples to **Universal Disagreement**. They are counterexamples because the psychological feasibility role strengthens, in the terminology above, the characteristic roles of practical terms. The psychological feasibility role is satisfied by a subset of the actions that the normative 'ought' or the moral 'wrong' might be applied to, and this additional role has the function of shifting

engage in a substantive disagreement with us about whether an act is morally wrong, even if they have a different view about what constitutes a legitimate excuse. The Feasibility Twins, by contrast, do not view psychologically infeasible acts as sometimes wrong, and always excused. Instead they view such acts as never wrong.

the property users of practical terms are talking about.

2.4.3 Overriding roles

Here is a final case. Take a Twin community that uses a practical term with the *bounded optimality role*. A term used with this role is applied in a pattern which tracks what some theorists have called "bounded optimality". At a first pass, bounded optimality can be characterized as follows: an action is boundedly optimal (or rational) when it is the action that the principles that it is best for an agent to encode, given their limitations, typical environment, and cognitive architecture, would recommend.[71]

There are a number of reasons to think that bounded optimality is a theoretically interesting notion. Most real-world agents face cognitive limitations that make taking the optimal action almost impossible: the number of possible moves in a chess game is finite, and there is in principle no information that is unavailable to a player about the outcome (win, lose, or draw) a possible tree leads. The knowledge is available, but no human has enough brainpower to compute and remember all but a very small number of possible moves (Simon, 1972, 169). In realistic circumstances one has to make moves in chess after entertaining only a small range of possible strategies.

In addition bounded rationality is relevant to agents for whom deliberation is costly. Processing information and evaluating options can take time and effort, both of which take energy and waste potentially valuable resources. (Sometimes taking time to figure out what to do can be very costly: for instance, when you need to figure out how to defuse a bomb on a timer.) Limited agents who need to make a decision might fail to make the optimal decision. In a sense, this is a failure: they are failing to choose

[71] See Simon (1972) for an early elaboration and Conlisk (1996) for an overview of debates over bounded rationality in the economics literature.

what is best. But this way of framing to problem points to another sense in which the agent can, in another sense, do what she should be doing in not choosing what is best: she is choosing the boundedly optimal action.

As agents we adapt to these limitations by adopting rules of thumb. These are relatively simple rules that are easy to compute with the cognitive architecture that we have. They are not shared with all other possible rational agents. Other possible agents can more easily process different information, or possess different cognitive architectures, that make it most efficient to process a different set of action-guiding rules. These rules sometimes recommend actions that are best. Having good rules of thumb to deliberate with makes sense both when deliberating to find the optimal action almost certainly won't lead us to act optimally (as in chess), or when aiming at optimality carries costs (as in defusing bombs).[72] Since a boundedly rational agent follows these rules of thumb, they will always do the boundedly optimal act—but will not always do the act that is best simpliciter.

A few additional notes about bounded rationality are in order:

First, the features of an agent that make her decision boundedly rational are not just additional features of a decision context that a theory of obligation should take into account. Perhaps I could checkmate my opponent in a limited number of moves if I move my rook, but noticing this would require exploring a strategy that is rarely successful, and difficult to for an agent like myself evaluate. Instead I move my queen, which only marginally improves my chances of winning but is recommended by a simple and familiar strategy. (We can suppose, for simplicity, that there are no other

[72]On one view, the rules a boundedly rational agent uses are satisficing rules (Simon, 1972), though I will not commit to any specific theses about the nature of these rules here. Even if these rules have the structure of satisficing rules, it is important to note that these will not be the same rules that a satisficing consequentialist normative theory, as discussed in Slote and Pettit (1984), would recommend.

moves available that would increase my chances of winning.) The circumstances surrounding my move that involve the decision-making process are not additional facts that, when considered, show that moving my queen is in fact the best move, simpliciter. Boundedly rational agents are not agents who are deciding what the obligatory move is, and are simply taking different facts—such as the fact that the strategy that recommends moving the rook is rarely successful—as relevant to conclusions about what is obligatory.

Second, there are analogues of optimal and bounded rationality that correspond to moral decision-making. Optimality simpliciter corresponds to doing the morally required thing. Just as what is optimal, or best, to do is a question for substantive normative theory, what is morally required is a question for substantive moral theory. Substantive moral theory might hold that doing the action that maximizes happiness is morally required, or that not acting in ways that violate the autonomy of rational agents is morally required. Doing what is morally required involves doing whatever the correct moral theory recommends.

A boundedly rational agent will not always be in a position to know what action is required in this sense. For the same reasons as before, an agent who follows good rules of thumb, given her limitations and cognitive architecture, will not always do the morally required action. Sometimes reasoning to the morally required action involves difficulties or costs for limited humans. There are rules of thumb for moral decision-making that are best for agents like us to employ.

With these clarifications about the bounded optimality role in place, we can imagine a Twin Earth case involving communities who differ over their use of practical terms with the bounded optimality role. Begin with a community of Earthlings who use their

normative term 'ought' to refer to obligation simpliciter. Sometimes, they allow, we will not be able to know what we ought to do because of limitations of information, time, and cognitive limitations. For instance, in a game of chess where an obscure move of the rook will increase one's chances of winning dramatically, even though no normal agent would be able to notice this during a game, speakers on Earth say 'one ought to move the rook'.

On Twin Earth, speakers use their term 'ought' with a normative role, but in addition use the term with the bounded optimality role. These are the *Bounded Optimality Twins*. Thus whenever there is an action that would not be required by the principles of bounded rationality, the Bounded Optimality Twins do not apply their normative 'ought' to it. When considering the same chess game as the Earthlings, the Twins say 'one ought not to move the rook'.

The Bounded Optimality Twins are using a primitive normative term. Even though they apply their term 'ought' only to acts that we would describe as acts that are required by the best rules of thumb for agents like the Twins to use in decision-making, the Twins do not conceptualize their use of their normative 'ought' in this way. They don't have another normative concept of what it is 'best' to do, and then define the application conditions for the 'ought' with the bounded optimality role in terms of this prior notion. Instead, they simply have a disposition to apply their term 'ought' to the action that in fact is boundedly optimal, and will view considerations of economy and cognitive architecture as relevant considerations when determining what 'ought' applies to. Since the disposition is, we can suppose, near-unanimous throughout the Twins' community and psychologically robust, it qualifies as a role, and appears to be a significant semantic constraint on the referent of their 'ought'.

The Twins assert what, on the surface, is a sentence that contradicts what speakers on Earth say. But the most natural interpretation of the case is that there is no genuine disagreement between Earthlings and the Bounded Optimality Twins. On Twin Earth, speakers are referring to the property of being boundedly optimal. So, when they assert 'one ought not move the rook', they are asserting that moving the rook would not be boundedly optimal. Meanwhile, on Earth, speakers are referring to obligation simpliciter. When they say 'one ought to move the rook', they are speaking about obligation, a different property. Each can agree with what the other is saying, in their own language: we can recognize that Twin Earthlings truly say that moving the rook is not boundedly optimal. Any appearance of disagreement is merely verbal.

All of this is compatible with each community using their 'ought' with the Gibbard role. When saying 'one ought not move the rook', the Bounded Optimality Twins will regard someone who agrees and moves the rook anyway as incoherent. Likewise when saying 'one ought to move the rook', speakers on Earth will regard someone who agrees and fails to move the rook as incoherent.

The bounded obligation role *overrides* the characteristic roles of practical terms. If the normative 'ought' in English refers to obligation, the actions that satisfy the bounded obligation role sometimes fail to be obligatory, and sometimes actions that don't satisfy the bounded obligation role are obligatory. (In the chess case, moving the queen is boundedly optimal and moving the rook is obligatory.) Use of a practical term with this role appears, in some possible cases, to refer to actions which have the property corresponding to the bounded obligation role, and not the property that best fits the normative role simpliciter.

According to **Universal Disagreement**, this kind of overriding cannot happen: any

community that uses a term with the same normative role refers to the same thing. But it seems clear that this is false. In addition to the overriding roles, it also appears that there are weakening and strengthening roles which shift the reference of practical terms. These are not the kinds of case that a typical presentation of the Moral Twin Earth case focuses on. But they are important for understanding the range of possible disagreement between users of practical terms. The usual Moral Twin Earth cases do show something interesting—namely, that practical terms exhibit what I am calling *robust* stability. Possible communities can use their practical terms in accordance with very different substantive normative theories, and still manage to talk about the same thing. But a tempting generalization of this conclusion—that the stability for practical terms is *universal* in a sense that supports **Universal Disagreement**—is false.

This is an important datum for any meta-semantic theory for practical terms. I will turn to a positive theory in subsequent chapters. But it is first worth remarking on some additional details of the counterexamples to **Universal Disagreement** listed here.

2.5 Possible disagreements with a contextualist semantics

Universal Disagreement is a claim about *substantive* disagreement. Substantive disagreements, as I have characterized them, require shared reference. This is what distinguishes a genuine disagreement from a verbal dispute: speakers are talking about the same thing, and not merely using words that have the logical form of a disagreement with different meanings.

The notion of a substantive dispute is not supposed to be a friendly characterization of the issue for those who prefer non-cognitivist modes of theorizing about

these issues. It may be that, if non-cognitivism is true, a substantive disagreement involving practical terms should be characterized differently. Disagreement in plan, in Gibbard's sense, may be a perfectly adequate notion of substantive disagreement for a non-cognitivist. The motivation for focusing on substantive disagreements is internal to the realist theory I am developing here. If the realist can explain the intuitive Moral Twin Earth-style disagreements as substantive, then there is no room for the objection that the realist's explanation of the phenomena is in some way incompatible with the core tenets of the realist view.

The realist holds that facts about obligation are part of reality (reality "favors certain ways of acting"), and practical language describes this aspect of reality. So a straightforward approach to implementing realist ideas involves the simple idea that, in a case of a substantive normative dispute, both speakers use their term 'ought' to refer to the same property—obligation—and make incompatible claims about it. Similarly for substantive disputes about morality.

This view of practical language makes an assumption, however, that is tangential to realism. This is the assumption that the central practical terms 'right', 'ought', and the like, are predicates which have the semantic function of referring to a property. The simple syntactic assumption is very natural (the sentence 'Jamie ought to take out the trash' appears to contain a 2-place predicate). Moreover, in many cases is a harmless simplifying assumption even if it is false. I will continue to use the language of reference in describing the semantic function of practical terms throughout this book. But it is worth making some remarks on how to understand these issues within the more sophisticated semantic framework. inspired by Kratzer (1977), as outlined in Chapter 1.

The contextualist view does not supply anything the can be described as "the referent" of a practical term. This is for several reasons. One is the simple reason that the theory is a contextualist theory: 'ought' in some contexts expresses a moral notion, by using an ordering source that ranks actions according to a moral standard. But this is not an invariant feature of 'ought': in other contexts it expresses a prudential notion, by using an ordering source that ranks actions according to how prudent they would be for an agent, given her desires and situation. Moral obligation cannot be said to be *the* referent of 'ought', for the simple reason that the term as used in some contexts is not about morality at all.

There is also a deeper issue, which remains even if we limit attention to only contexts where 'ought' is used as a moral notion.[73] Since 'ought' in these contexts applies to an action which, out of a contextually relevant set of alternatives (the "modal base") ranks highest according to relevant moral standards (the contextually salient "ordering source"), it is forced at best to speak of a *property* that all of these moral uses are referring to.

The primary reason is that 'ought' is best thought of as an *operator* on the contextualist view. If Φ is a well-formed sentence, then 'Ought Φ' is as well.[74] The sentence is true just in case Φ is true at the (contextually salient) worlds that rank highest accord-

[73]What these contexts amount to is not entirely clear. I have been following a tradition according to which what is distinctive of moral terms is their role, and in particular their role-like connections to blame and guilt. A context-sensitive 'ought' does not invariably have this role. Instead at best 'ought' *in particular contexts* can be said to be used with the moral obligation role. The connection between a use-in-a-context and this role is not obvious. A role, by definition, is a community-wide feature; 'bachelor' has a particular role in English because English speakers as a whole use it in certain ways (nearly unanimously treating 'bachelor' as applying only to the unmarried, etc.) But a use-in-a-context is a feature of an individual: at noon on Sunday, Ellie might say 'we ought to leave now'. What makes this particular use of 'ought' a moral ought, i.e., a use of 'ought' with the moral obligation role? I will not explore this question in detail, but a plausible answer in schematic form is that there is a community-wide precedent of using 'ought' with the moral obligation role in some contexts. At noon, when Ellie tokens a sentence with the term 'ought', since intends to use 'ought' in accordance with that precedent, and (perhaps) her audience can tell that she has this intention. Community-wide usage makes other precedents available as well, but in this context Ellie does not intend to connect her use with these roles.

[74]Broome (1999) and Wedgwood (2006) adopt views about the syntax of 'ought' along these lines.

ing to the (contextually determined) ordering source. The semantic function of 'ought' is, in other words, to shift the world of evaluation.[75]

What would a substantive disagreement between communities that use 'ought' with a Kratzer-style semantics look like? Suppose the modally separate deontologist and consequentialist communities in the original Moral Twin Earth case are using the term 'ought' with the semantics as laid out by Kratzer. If they are using the same contextually supplied ordering source and modal base, then the communities in this case can be making the kinds of claims that generate a substantive disagreement. But there is no substantive disagreement if, for example, one speaker is claiming that giving money to charity ranks highest against the standards of morality, while a second speaker claims that giving to charity does not rank highest in view of what would best satisfy the donator's desires.[76] So we should say that the consequentialists on Earth and the deontologists on Twin Earth are having a substantive disagreement about the moral status of ϕ-ing only if one of the communities claims that ϕ-ing ranks highest against a set of available moral standards, and the other community claims that ϕ-ing does not rank highest against the *same* moral standards. Reference to a property of obligation has no role in explaining substantive disagreements on this picture, but there is still a notion of substantive disagreement that is available.

Here is a simple case to illustrate this idea. Suppose an Earthling is deciding among the following two actions (and does not take there to be any alternatives): taking her child to a movie, and giving the money the movie would have cost to a charity that

[75]See Silk (2017) and Chrisman (2015). Chrisman argues that the Kratzer framework will not capture some of the central claims of this book which involve so-called "agential" oughts. See Dunaway (2017a) for discussion.

[76]Of course in contexts where these are the relevant standards, speakers will say 'one ought to give money to charity' and 'one ought not to give money to charity', respectively. A similar point can be made about contexts where different modal bases are in play.

provides famine relief. She wants to decide on an action for purely moral reasons, and so does not take any other standards to be relevant. As a consequentialist she decides to give the money away: the money will potentially save a life when given to the charity, and will only produce a small amount of temporary happiness for the child. She expresses her judgment by saying 'I morally ought to give the money to charity'. This claim is true just in case the relevant moral standards rank giving the money to charity over taking the child to the movie.

A Twin Earthling might be engaged in making the exact same decision, considering only two actions (taking her child to the movie or giving the money to charity) and wanting to make a decision based only on moral considerations. As a deontologist, she thinks she has a duty to sometimes benefit her child, and so says 'it is not the case that I morally ought to give the money to charity'. This is true, given the Kratzer semantics, just in case the relevant moral standards do not rank giving the money to charity over taking the child to the movie.

The Earthling and Twin Earthling are having a substantive disagreement if the "relevant moral standard" is the same for each speaker. If the moral standards in each case are distinct, then the claims they are making are not incompatible, and so no substantive disagreement is possible. If the rankings are distinct, then the speakers might both be speaking truly—and so a disagreement that is substantive is not possible.[77]

It is obvious that each speaker *thinks* that the moral standard they accept is the standard that determines the ranking at issue for the truth-conditions of their own moral claims. On Earth, speakers take the relevant moral standards to be consequentialist standards, ranking actions according to how much happiness (or some other combina-

[77]It is worth emphasizing again that non-substantive disagreements are still possible. See Silk (2017) for one account of how a contextualist framework might be put to use in giving an account of non-substantive disagreement.

tion of goods) they produce. On Twin Earth, speakers are not consequentialists, and so think that the relevant moral standards sometimes fail to give the highest ranking to actions that produce the most happiness. The crucial point for a realist explanation of the case as an instance of substantive disagreement is that it needn't be that each community is *right* about what the moral standards are. If they were, then the disagreement between the Earthling and Twin Earthling would not be substantive, but it is not essential to the contextualist picture that speakers are authoritative over what ranking is at issue in this way.

It might be tempting to think that Kratzer semantics does entail that speaker intentions to use a particular ranking determine which actions rank highest in the relevant context. A model that takes the context-sensitivity in 'ought' as closely analogous to other context-sensitive expressions will naturally have the implication that speakers on Earth and Twin Earth are talking past each other. For example it is relatively natural to interpret a parent who sees that all of his children are in the car as speaking truly (in his context) when he says 'everyone is in the car'. Since a parent's interest in a case like this usually involves ensuring that no children are left behind when the car drives away, the parent's intention to include only his own children in the domain of his quantifier 'everyone', quite naturally, determines that in this context it is only the location of the relevant children that matters for the truth of the sentence 'everyone is in the car'. By contrast a nearby police officer in the same situation would naturally reject the sentence 'everyone is in the car'. The officer is concerned with the safety of the general public and so would treat other nearby people who are not in the relevant family as in the domain of his quantifier. So the officer would speak falsely if in

asserting 'everyone is in the car' in the relevant context.[78]

Treating the intentions of the Earthlings and Twin Earthlings analogously threatens to fail to predict any substantive disagreement on the contextualist semantics. But there is no reason to take the analogy this far. We should first distinguish between an intention to use 'ought' as *as a moral term* from an intention to use 'ought' *with a ranking that is implied by a particular moral theory*. The first is the "flavor" of the ranking at issue; when 'ought' is used in a context to reflect moral standards, it has a moral flavor. In other contexts the flavor is not moral. Whether a particular use of 'ought' has a moral flavor plausibly is determined by speakers' intentions. If Sally says 'you ought to give some money to charity' and intends to speak about what morality requires, then it is very natural to interpret her 'ought' as having a moral flavor on that basis.

The moral flavor of a use of 'ought' should be distinguished from ranking actions in accordance with a particular theory. Suppose Sally says 'you ought to give some money to charity', and does so on the grounds that giving money ranks highest according to a particular consequentialist theory of morality which she accepts. For specificity, let's suppose that the particular theory is one that ranks actions according to how much happiness they produce; we can call the ranking in question the *happiness ranking*. It is not required by a contextualist semantics that, simply because Sally accepts a form of consequentialism that entails the actions highest on the happiness ranking ought to be done, her assertion is true just in case giving some money to charity ranks highest on the happiness ranking.

The contextualist account specifies the parameters that context fills in for 'ought' and related expressions. It does not specify in great detail *how* these parameters are

[78]Though this does not preclude a kind of disagreement between the parent and the police officer, namely disagreement over which individuals should be in the range of the quantifier. See Silk (2016).

filled in, in a given context. Keeping the idea that there are substantive disagreements between communities in a Moral Twin Earth scenario, along with the Kratzer semantics, places some significant restrictions on how these parameters are filled in. A realist will resist the idea that a speaker's intention to rank actions according to a particular ranking determines a ranking in a context. Instead it will need to hold that, in general, speakers who use the same flavor of 'ought' are thereby capable of making incompatible claims with their use of 'ought', since the shared flavor implies shared ranking. If the speaker intends to use 'ought' with a moral flavor, then the truth-conditions of her assertions are determined by a ranking of actions according to moral standards. But which actions rank highest according to these standards is determined by the objective facts about morality, and not what the speaker thinks about the content of the moral standards.

We can summarize this thesis in the following claim:

Ranking Stability Any possible speakers who use their term 'ought' with the same moral or normative flavor are thereby making claims whose truth conditions are determined by a single ordering source.

Ranking Stability mimics the kind of referential stability that **Universal Stability** says is required for substantive disagreement on a simple referential semantics for practical terms. In cases where the simple view would say that two speakers are referring to the same thing—a measure of stability for practical terms—the realistically-inclined contextualist account should say that speakers are making claims with a term that has its truth-conditions determined by the same ordering source. Any substantive disagreement arises because they are making incompatible claims about which actions ranking highest according to the relevant ordering.

Of course, the same considerations which show that **Universal Stability** is false for a simple semantics will also require refinements for **Ranking Stability** on the contextualist semantics. Given that a term used with both the moral rightness role and the bounded optimality role will not be referring to the same property as a term used with only the moral rightness role on the simple view, the refinement to **Ranking Stability** will need to account for the possibility that two speakers who are using 'ought' with a moral rightness role are not making claims about a single ranking on the contextualist view.

The contextualist account will raise complications for the positive picture I present in Chapters 3 and 4. I have argued here that it does not present any in-principle obstacle to the points I have made so far. For simplicity, I will continue to conduct the primary discussion in terms of a simple referential semantics, and not the Kratzer-style contextualist semantics. It will, however, be instructive to return to the view at various points in what follows.

2.6 Lessons and the way forward

We should reject **Universal Stability** as an explanatory desideratum for a meta-semantic theory for practical terms. This does not mean that there is nothing to explain.

A meta-semantic theory for practical terms will need to explain why they are *robustly* stable. This requires explaining why communities in the original Moral Twin Earth scenario and nearby variants are referring to the same thing, and are thereby capable of having a substantive disagreement. The differences in use of practical terms between these communities can be strikingly large. The appearances are that large components of usage of practical terms by a community, aside from normative or

moral role, are irrelevant to what they are talking about.

But a meta-semantic theory for practical terms will not entail that the stability is *universal*. It should not entail that every community which uses their practical terms with the same role is thereby talking about the same thing. We have examples of possible communities who use terms with a shared practical role, but appear not to be talking about the same thing because they do not disagree. I have given some examples, including the Ability Twins, the Feasibility Twins, and the Bounded Optimality Twins, but there may be other examples as well. The lesson is that not every additional role must made a difference to reference. But some clearly do. A term's normative or moral role is not necessarily the only aspect of use that determines reference.

If these claims are correct, then they highlight an important distinction in the semantics of practical terms. On the one hand, there is the semantic contribution of a practical role. Take the Gibbard role: if a community has a normative term, then they systematically use it as a term that closes off deliberation, since they think someone who applies the term to an act and does not do it is making a kind of conceptual mistake. It is plausible that language of this kind is indispensable—any agent will need a way of settling what to do, in a way that is characteristic of normative terms. But not every property that attaches to the semantically relevant role of normative terms will thereby attach to the *referent* of normative terms. In particular, the indispensability of normative language does not imply that reference to a particular normative property is indispensable. Even if every possible linguistic community must use a term with a normative role, it does not follow that they must use it to refer to the same property. The Ability Twins and others provide examples of possible linguistic communities that do not refer to normative obligation. But they still have a term that plays the relevant

deliberative role in their language.[79]

A second, related point is that, in light of the failures of **Universal Stability** there is a further question about what language it is best to speak. Nothing I have said rules out the claim that the Ability Twins, or other possible communities who are counterexamples to **Universal Stability**, are making a mistake of some kind. It cannot be a mistake that these possible communities will be able to express using their own practical terms. Nor is it is a mistake that will invariably show up in the form of contradictory beliefs, which would be an especially glaring defect. But the possibility remains that even if they use 'ought' in a way that makes the sentence ⌜one ought to ϕ in circumstance c⌝ true in their language, where ϕ is an action that is not obligatory in c, they are making a mistake of using a term that has these truth-conditions. (Perhaps the mistake is *ineffable*, in the sense of Eklund (2017, Ch. 2).) So far I have only described the semantic profile of practical terms; the significance for important questions about the nature of normativity I will defer until the conclusion.

The question I will pursue in Chapters 3-5 is: what does a realist explanation of these phenomena look like? I will not be asking what a non-cognitivist should say about them,[80] nor will I be interested in responses that try to explain away the data rather than explain it.[81] The primary question I will be interested in is instead whether realist assumptions can explain why reference and disagreement with practical terms should pattern in this way.

The picture I will paint sounds, for the most part, a positive note. I will also identify some limitations of the realist approach to the data. The theory I present aims not only

[79] I thank a reader for raising this distinction, and highlighting its significance for me.
[80] The failure of **Universal Disagreement** is an interesting question for the non-cognitivist, though not one I will pursue here.
[81] Though some of what I say in subsequent chapters can be adapted to this approach. See Dunaway and McPherson (2016) and Manley (MS) for details.

to account for the data, but to explain it as well—that is, it provides a plausible story about why we should expect to see disagreements and their absence where we do. This is not by itself an argument for realism. But it is a part of a development of the view that has certain attractions. It shows that, once we accurately characterize the data about disagreement with practical terms, the realist has a plausible explanation of the phenomena. It also provides a challenge to competing views to provide a similarly adequate explanation.

Chapter 3.

Reference Magnets: How Do They Work?

The first two chapters characterize a target for the realist to explain. The target is not **Universal Disagreement**, which is false. Instead, it should explain a more limited claim: that practical terms are *robustly* stable, and support disagreements between a wide range of possible linguistic communities. **Robust Disagreement** provides one characterization of the explanatory target. Some possible communities who use their practical terms in very different ways still manage to refer to the same property. This is a striking explanatory desideratum; many descriptive terms are not so stable. Possible linguistic communities that apply 'red' to different color-shades are not talking about redness, and even a natural kind term like 'water' will refer to something besides H_2O in different environments.

But not every possible community that uses a term with a practical role will refer to the same property. Some communities use practical terms with the same moral or normative role, but fail to have substantive disagreements, since they are referring to different properties. Practical roles do not determine reference, but they do beget significant semantic stability.

If moral and normative properties are a part of reality, as the realist holds, then there needs to be an explanation of how these properties end up as the semantic content of our practical terms. It is not just that we do manage to refer to these properties with our terms 'right', 'ought', 'should, and the like. It is that speakers in many other possible communities, some of whom have very different moral and normative views, and use their practical terms accordingly, *also* manage to refer to the same properties. The explanation for this cannot be simply that every community that uses their 'ought'

with the same normative role refers to the property of being obligatory—this entails **Universal Stability** and **Universal Disagreement**. These theses are too strong. We need an account that does the twin jobs of explaining how many possible communities refer to the same part of reality with their terms that share a practical role, while at the same time placing the right limits on how far this phenomenon of shared reference extends.

The rest of this book develops one explanation on behalf of the realist. It has a metaphysical component: the properties that practical terms refer to are metaphysically elite, and there are multiple highly elite properties that practical terms can refer to. It has a linguistic component: that practical terms refer to properties on the basis of how they are used and how elite the candidate referents are. And it has an epistemological component: it predicts the contours of which possible communities disagree with each other, and which do not, on the basis of how we can know where the elite moral and normative properties are.

This chapter and the next develop the metaphysical and linguistic components. I begin in this chapter by defending the claims that there are some metaphysically elite properties, and that it is relatively easy for language users to refer to these properties. This package of views is sometimes called *reference magnetism*. Reference magnetism is a thesis about metaphysics and language in general, and is not specific to the properties of rightness or obligation, or the semantics of practical terms. Later chapters will apply reference magnetism to this specific case. But the principles behind it are claims about metaphysics and language that are not tailored specifically to solve problems in metaethics. I develop these claims in this chapter. The principles are general. Their role in explaining the distinctive features of practical language, as I argue in later chapters, is

a consequence of how these principles apply to practical subjects.

3.1 Magnetism: an introduction

David Lewis introduced the thesis that has come to be known as reference magnetism in a series of papers[82]. The concept has been taken up in a wide range of recent theorizing[83]—though the various parts of the implementation may not all have been claims Lewis would endorse.[84] Although many of the details are controversial, it is a powerful theory that is also quite simple.

In this chapter I motivate, develop, and defend reference magnetism. I first sketch its role as a theory in meta-semantics, and then outline its two key components. These are the notion of an elite property, and the idea that eliteness is one important factor in what determines reference. Since most discussions of reference magnetism in the literature are critical, and there is little existing literature that defends the basic view from these objections, I also provide a defense of the view from some recent criticisms. These defenses will be instructive in developing the view further, and set the stage for an application to practical terms in an explanation of why **Robust Disagreement** holds but **Universal Disagreement** fails, in Chapter 4.

3.1.1 Use and meta-semantics

Reference magnetism is, in the first instance, a thesis within *meta-semantic* theory. A meta-semantic theory explains why our language refers to what it does—its subject matter. Part of the explanation will necessarily involve certain facts about our linguistic activity. The words 'dog' and 'Chevrolet' are about different things, and this is because

[82] Lewis (1983, 1984)
[83] See Weatherson (2003), van Roojen (2006), Sider (2012), and Dorr and Hawthorne (2013) for some examples.
[84] Schwarz (2014)

we use the terms differently. But, differences in how we use terms is not sufficient to explain why we do refer to dogs and not any one of a host of possible referents in the neighborhood, including the candidate referents dogs-before-3000 AD and dogs-that-aren't-owned-by-Cicero. Our linguistic activity alone will not rule out all of these less-privileged candidates. Something outside of our linguistic activity must feature in any good explanation of how our language acquires its subject matter.

Reference magnetism is the claim that reference is determined by not only how we *use* our language, but also by how eligible the various candidates for reference are. A first-pass motivation for this is that use, on its own, does little to single out a precise referent even for simple, ordinary terms. Given how we use the word 'dog', dogs-before-3000 AD; dogs-that-aren't-owned-by-Cicero; dogs-minus-the-molecules-touching-a-liver; dogs-not-under-a-red-umbrella are all candidate referents, since they fit pretty well with how we use 'dog'. Strictly speaking, we have probably successfully used the word 'dog' in a way that makes the referent dogs-before-3000 AD a poor fit because we say things like 'barring an apocalyptic event, dogs will continue to exist after 3000 AD'. Still, there will be other very similar candidates that we do not rule out; some of these candidate referents are so complicated and bizarre that it hasn't occurred even to philosophers writing on meta-semantics to list them. Presumably our word 'dog' doesn't refer to these even more bizarre candidates. But we can't plausibly say, of each one, that our usage determinately rules them out.[85]

Examples help to illustrate the point, but the problem has a very general source. Language use is limited, as there are finite instances of a tokening of the English word 'dog'. But the candidate referents for 'dog' are unlimited. There are an infinite number

[85]This is one of the issues connected to the "rule-following problem" discussed in Kripke (1982). Quine (1960) and Putnam (1981) raise similar issues.

of bizarre referents that need to be ruled out,[86] and it is not at all clear how use alone can do this. We can make general statements using 'dogs': 'dogs have teeth' seems to rule out any non-teeth-having things as the referent of 'dogs'. But even here as Putnam (1981) has pointed out that this only works if the other terms in the generalization (e.g., 'teeth') have already have determinate and non-gerrymandered referents. This is a significant motivation to include something more than mere use in our theory of what determines reference.[87] Reference magnetism is a meta-semantic theory which holds that the extra ingredient is provided by the world: certain candidate referents are more eligible in virtue of their metaphysical status.

3.1.2 Precisification and overridingness

There are two contributions for magnetism in a theory of reference-determination, which are in principle separable.

The first is *precisification*. Language use is, on its own, not enough to rule out a host of bizarre and gerrymandered referents that nonetheless fit with the usage. For example, if we haven't made the effort to use 'dog' to explicitly apply to a hypothetical dog owned by Cicero, it doesn't follow that the referent of 'dog' is indeterminate, failing to discriminate between dogs and dogs-that-aren't-owned-by-Cicero. The latter property is determinately not the referent of 'dog'. Use alone does not do enough to ensure that 'dog' is not indeterminate in reference, in the sense that use alone fails to rule out a number of distinct but nearby candidates. Reference magnetism is needed to rule out the gerrymandered candidates, and thereby precisify reference.

Second, magnetism *overrides* use in some cases. Sometimes we apply terms mistak-

[86] Cf. Lewis (1983, 346); also the "bubble puzzle" in Williams (forthcoming, §1.3).
[87] Similar points apply if we extend the list of reference-determining features to include *dispositions* to use our terms.

enly. A community of speakers of a language like English might sincerely insist that 'dog' does not apply to Pomeranians. Nonetheless, they would speak truly if they were to assert the sentence 'Pomeranians are dogs'. This is so even if a range of speakers are confused about Pomeranians. Even if, for these speakers, property of being a non-Pomeranian dog is a better fit with their use of 'dog' than dogs are, these speakers can be making mistakes and do not automatically speak about something different than what English speakers use 'dog' to refer to. Magnetism can override these mistakes.[88]

3.1.3 Elite properties: a schematic characterization

Why is it that 'dog' is not indeterminate between dogs, dogs-before-3000 AD or dogs-that-aren't-owned-by-Cicero? And why does it refer to dogs even for a possible community that uses 'dogs' to fit better with non-Pomeranian-dogs? The core idea behind reference magnetism is that doghood is a property (or entity[89]) that is distinguished metaphysically from other candidate referents. It is distinguished because, in Lewis's language, it is "more natural" than the alternatives since it "carves nature at the joints".[90] Even though a dog owned by Cicero and a dog owned by me differ in one sense—one has the property of being a dog-that-isn't-owned-by-Cicero whereas the other doesn't—this does not mark a metaphysically significant or genuine difference between the two. The difference between Cicero's dog and mine on this count is not, for example, a substantial difference in the way the difference between a dog and a giraffe is.[91] I will capture this idea by saying, following Lewis, that doghood is a

[88] These are cases that reveal a degree of semantic stability, in the sense of Chapters 1 and 2, in the natural kind term 'dog'.
[89] Hirsch (1997) treats magnetism for properties and entities separately. I will elide the distinction here.
[90] Lewis (1983, 346-7)
[91] Compare Sider (2012, Ch. 1) on 'Structure'.

more *elite* property than a the property of being a dog-that-isn't-owned-by-Cicero.[92]

Elite properties are unified and not gerrymandered. The unity in question is not a feature of the language we use to to talk about them, or the way we happen to think about them. Speaking loosely, we can say that eliteness is mind-independent: even if there were no minds, some properties (doghood) would be more elite than others (dogs-not-owned-by-Cicero). These are objective, metaphysical features of the properties in question.[93]

The (roughly labeled) mind-independence of eliteness is crucial for the role it plays in meta-semantic theories that incorporate reference magnetism. A meta-semantic theory, as I introduced it earlier, aims to explain why terms in a language refer to what they do. The elements of such a theory need, in other words, to be semantically independent features of the world. This is part of the explanatory nature of the project: if the meta-semantic theory used semantic facts about what words mean, or what speakers are thinking, then it would make use of semantic facts in explaining what terms refer to. A theory that relies on a linguistically or mentally constituted property would not be explanatory, since it would appeal to facts of the kind it is supposed to explain.

The fact that metaphysically elite properties are unified, non-gerrymandered properties is also important. Simplicity in a theory is a virtue. What simplicity amounts to is a difficult question, but it seems clear that explanations that involve non-gerrymandered properties are simpler than, and so are to be preferred over, explana-

[92] I will use terminology that differs from Lewis's because his use of the word 'natural' can be easily confused with other distinctions. One is the distinction between natural and *super*natural properties: for instance the difference between the biological species *homo sapiens* and the theological category *cherubic angel*. Another is the distinction between natural and *non*-natural properties, as used by Moore (1903) and the subsequent meta-ethics literature. It is, in rough outline, possible for a property to be natural in Moore's sense but very unnatural in Lewis's.

[93] Sider (2012, §4.6)

tions that involve gerrymandered properties.[94] Reference magnetism can be motivated by the idea that a preference for simplicity should extend to explanations of semantic facts.[95] Since a meta-semantic theory is explanatory—it offers an explanation of what terms in a language mean—good meta-semantic theories will on balance treat an interpretation of a language as better to the extent that it interprets speakers as referring metaphysically elite properties. Any theory of a language that assigns elite referents will on that basis be simpler than a theory that assigns gerrymandered referents.[96]

3.2 A (partial) theory

So far I have sketched some motivations for reference magnetism. Some are criteria of adequacy: a good semantic theory will not imply that 'dog' is indeterminate. There is the possibility that speakers are mistaken. Other motivations are methodological, including the connection between eliteness and theoretical simplicity.

These are, however, merely gestures at a complete theory. In order to assess the full case for reference magnetism, we will need a concrete theory on the table. In spelling out a concrete theory, I do not mean to endorse, or defend, the theory in every detail. Instead, the initial presentation of the theory will be deliberately simple. I will add some plausible additions to it while defending it from objectors below. Even so, it will remain neutral on a lot of questions within meta-semantics. This is intentional. The aim of this section is not to provide a complete meta-semantic

[94]Manley (forthcoming) provides an alternative on which simplicity is a theoretical primitive. Here I will not take a stance on whether simplicity or eliteness is explanatorily prior; the only point that matters for present purposes is that, in general, a theory that references elite properties will be simpler than one that does not. See also Lewis (1994).

[95]Williams (2007), Sider (2012)

[96]Simplicity is not the only theoretical virtue, and so the preference for elite referents is not absolute. A good theory may assign non-elite references in order to maximize other theoretical virtues. Also, simplicity is a holistic virtue of a theory—one part may be complicated, in order to achieve reduction in complexity elsewhere.

theory;[97] rather it is to sketch a theory in enough detail that the contribution of a preference for metaphysically elite referents is apparent.

A simple version of reference magnetism appeals to two factors in reference-determination: *use* and *magnetism*. The central meta-semantic claim of this simple version is that reference is determined by maximization of these two components. This is the **Magnetism** claim:

Magnetism The correct assignment of referents for a language L is the one that best maximizes fit with use of L and eliteness of referents it assigns to terms in L.

Some preliminary notes about **Magnetism** are in order:

1. Fit with use and eliteness are degreed notions. A candidate referent can fit perfectly with a community's use of a term, or it can fit pretty well, or it can fit poorly. A notion of degreed eliteness is possible as well: some properties are perfectly elite; others are close to elite, others are not elite at all.[98]

2. Maximization is a flexible notion, as the answer to which theory does the best job on these twin criteria will depend on how these are weighted. I will revisit this component while developing **Magnetism** in response to objections later in this chapter; for now I will work with the informal assumption that each receives substantial (non-0) weight.

3. The contributions of use and eliteness to reference-determination are holistic. This is why the **Magnetism** claim is a condition on a correct assignments of referents to a whole language instead of a correct assignment of a referent to an individual term.

[97] This involves settling a large range of questions about what meta-semantics is, and how to proceed methodologically. See Williams (forthcoming) for one recent attempt. I will engage with some particular claims Williams makes about magnetism and normative terms later.
[98] See Chapter 4 for more on this possibility.

We cannot simply ask whether, given the use of 'dog' plus the eliteness of a candidate referent D, 'dog' refers to D. Rather, the referent of 'dog' can only be settled by a theory that assigns referents to each term of the language to which 'dog' belongs. An assignment of a highly elite referent that does extremely well in fitting the use of 'dog' might force any interpretation that includes it to assign not-very-elite and poorly fitting referents to other, related terms. In this case a high degree of satisfaction of the individual components of reference-determination would not be a good guide to reference. Officially, we need to treat reference-determination holistically.

With these notes out of the way, we can say more about the specifics of use and eliteness in **Magnetism**.

3.2.1 Use

Broadly, a community's use of a term is constituted by their dispositions to use that term in various circumstances.

It is not necessary that all aspects of use count equally toward settling reference. Begin with one aspect of use, which involves what we can call *individual applications* of a term. Normal English speakers will apply their term 'dog' to particular dogs at a time: for instance, I can apply, on Tuesday at 3 30 pm, the term 'dog' to Daisy, the dog in my house. Assuming things go normally, I will continue to apply the term to Daisy at later times. But individual applications of the term 'dog' needn't be entirely reliable; for instance some English speakers will refuse to apply the term to specific Pomeranians.

This does not automatically make doghood a poor fit with our use of 'dog', however, since individual applications do not exhaust usage. Usage is also encoded in

the *generalizations* speakers accept. For instance, English speakers will typically accept the sentence 'fruits are the edible structures of seed-bearing plants'. Accepting these sentences does not involve any individual application of the term 'fruit': a speaker may utter the generalizations without any particular fruits in mind. But they constitute part of the use of 'fruit' by English speakers. Tomatoes fit decently well with this use of 'fruit'—mistakes in individual applications of 'not a fruit' to tomatoes notwithstanding—in part because we accept the generalization and are disposed, under the right evidential circumstances, to grant that that 'edible structures of a seed-bearing plant' applies to tomatoes.[99]

There are also *compositional* rules determining reference for complex expressions.[100] 'Small dog' is an expression made up out of simple linguistic items. Thus the reference of this expression is not straightforwardly determine by use (and other factors, like eliteness) in the manner described by **Magnetism**. ('Small dogs' doesn't refer to small dogs simply because small dogs maximize fit with use an eliteness.) Instead reference-determination applies at the level of simple expressions first (though, in light of the role of generalizations in our use of simple expressions, part of their use will involve their appearance in complex expressions). The referent of 'small dogs' is determined, in the first instance, by the application of **Magnetism** to 'small' and 'dogs'. Compositional rules for the language will then determine how the complex expression 'small dogs' refers on the basis of the reference of its constituent simple expressions.[101]

[99]Some treat the 'theoretical role' or 'conceptual role' of a term as the sole determinant of the reference of terms: see Lewis (1970) on for this approach to theoretical terms and Wedgwood (2001) to normative terms. We can views like these as the limiting case on the degree of relative importance of generalizations and individual applications. However I will not wade into this issue, since even right theoretical roles will not solve the worries magnetism is needed for; see Hawthorne (1994).

[100]*Cf.* Lewis (1992), and discussion in Weatherson (2012).

[101]There is a another aspect of use that is worth accounting for: *deference*. Some speakers treat others in their community as having more knowledge, experience, or other kinds of expertise with respect to some subject-matters. We can call these speakers "experts"; in some cases a community at large will defer to experts' use of certain terms, treating the expert usage as an important (or perhaps the only) determinant of

How individual applications, generalizations, and uses in composite expressions combine to determine which candidate referent best fits the use of a linguistic community is a complex question. Here I mention them simply to note that it is an option for a theory of reference-determination to privilege uses of a term in generalizations, plus compositional rules, for determining the referents of complex expressions in a theory of reference-determination. At various points below I will exploit this option in developing and defending reference magnetism.

3.2.2 Eliteness

So far I have focused on these use component of **Magnetism**. Use is a feature of a linguistic community. But according to **Magnetism**, the correct theory of reference for a language also depends on features of the environment of speakers of the language: reference is biased toward elite, non-gerrymandered properties. Here it is worth characterizing what the eliteness of a property amounts to, in more detail.

For reasons that will be made clear below, I will begin be characterizing the notion of an elite property in terms of the *role* that it plays in relation to other theoretically interesting notions. The first-pass gloss on eliteness is that it distinguishes those properties which are not gerrymandered or disunified, from those which are. In the former category are properties such as: being an electron, being a number, and occupying spacetime. Gerrymandered properties include properties such as being a dog before 3000 AD, reflecting being either red-orange or blue-green but not red or blue, and being a number that is divisible by 3 and less than 5,429. Properties in the former

the reference of the term. For example English speakers plausibly defer with high-level physical terms like 'electron': ordinary speakers do not spend much time in cloud chambers and so have very little opportunity to engage in individual applications of 'electron'. Not many will accept general principles about electrons either. Deference to experts is the most plausible mechanism by which ordinary speakers manage to talk about electrons. This is illustrated by the "semantic division of labor" in Putnam (1973).

category are the kinds of properties that are relatively easy to refer to. These are the elite properties. But what they have in common is more than simply their eligibility for reference, and this is what makes them elite. Here are some especially significant roles that eliteness plays:

Similarity. If two things are electrons, they thereby resemble, or are similar to each other, to a significant degree. Conversely, if one is an electron and the other is not, they thereby *fail* to resemble each other.[102] Shared elite properties are similarity-conferring, and failing to share elite properties is dissimilarity-conferring. The same does not hold for non-elite properties. If two things are both either red-orange or blue-green but not red or blue, they might look nothing alike: one could be red-orange and the other blue-green. Likewise if Daisy and Fido differ in that Daisy is a dog before 3000 AD and Fido is not, it does not follow that they are very dissimilar. Dogs existing before 3000 AD need not be substantially different from dogs existing after.

Lawhood. It is common to make a distinction between merely true general generalizations, and genuine *laws*, or law-like generalizations. Even if it is true that a significant tectonic plate shift has never occurred on 11 53 am of the first Tuesday of an odd-numbered month, this generalization does not capture a genuine law-like generalization about earthquakes. A plausible reason is that the property of occurring on 11 53 am of the first Tuesday of an odd-numbered month is not very elite. In order to appear in a law-like generalization, we need a property that is highly elite.[103]

Induction and projectability. Puzzles related to the "new riddle of induction" in Goodman (1955) point to a role for elite properties in epistemology. An observation of

[102] Dorr and Hawthorne (2013, 21-25), Sider (2012, 63-64)

[103] Dorr and Hawthorne (2013, 20-21), Sider (2012, 21-23). As Sider emphasizes, there may not be a single central notion of lawhood. It is, however, much more plausible that there is a distinction between merely accidentally true generalizations and law-like generalizations, and it is this distinction that eliteness, as I conceive of it, is connected to.

negatively charged things repelling each other supports the belief in the generalization *negatively charged things repel each other*. But if we use Goodman's term 'grue' to refer to the property of being green and first observed before the year 3000 AD or blue and first observed after the year 3000 AD, then an observation of a grue emerald does not support the generalization *emeralds are grue*. Only fairly elite properties are *projectable*.[104]

Eliteness, as I will conceive of it, is the property that confers similarity on its bearers, contributes to the law-like character of certain generalizations, and makes some properties projectable. This is the *eliteness role*. There may be additional roles that eliteness plays: see Dorr and Hawthorne (2013) for an extensive list. But for starters, we can understand the notion of an elite property through its connections to similarity, laws, and projectability. **Magnetism** captures the idea that the properties that play the eliteness role are also highly eligible for reference.

It is worth contrasting this notion of eliteness with the claims Lewis made about it in his early papers that introduced reference magnetism. While the application of reference magnetism I will develop here owes much to Lewis, it is not intended to be a perfect replica of his theory, and in fact makes substantial departures on some important points. In Lewis (1983), Lewis holds that the property that plays the eliteness-role as I characterized it above is a very disjunctive kind. It is split into two metaphysically very different properties: the first is *perfect* eliteness, and the second is *less-than-perfect* or *degreed* eliteness. The basic idea is that some properties are perfectly elite, while the less-than perfectly elite properties are those that can be defined with comparatively short definitions that employ only terms for perfectly natural properties, plus logical

[104]Sider (2012, 35-38), Lewis (1983). Williams (forthcoming, Ch. 3) raises some issues about the details about the implementation of this idea.

connectives.

Here is an example. Suppose (with Lewis) that the perfectly elite properties are physical properties such as mass and charge. Doghood is not a perfectly elite property. But it is more elite than dogs-before-3000 AD, since (Lewis conjectures) the definition of doghood in terms that stand for perfectly elite properties plus logical connectives is shorter than the corresponding definition of dogs-before-3000 AD. The length of definition of doghood explains why it is more elite than being a dog before 3000 AD. But there is nothing in Lewis's view which explains why mass and charge are perfectly elite while other properties are not.[105] In general, then, there are two elements to Lewis's view of eliteness:

- The perfectly elite properties, which are *primitively* elite;
- The less-than-perfectly elite properties, which are explained in terms of length of definition in perfectly elite terms.

At the methodological level, it is not obvious that the definition of degrees of eliteness is necessary. Since Lewis offers no explanation of perfect eliteness, he takes the notion as a metaphysical primitive which we are justified in positing because it is theoretically fruitful. There is then no requirement that he not take eliteness to other degrees as primitive as well, on account of a similar theoretical utility. Perhaps there is some preference from considerations of theoretical economy for providing a definition if one is available. But note the definitional direction could go the other way around. If degrees of less-than-perfect eliteness are primitive, then perfect eliteness could be defined in terms of not having anything be more elite.

[105] Lewis (1983, 347), Lewis (1986, 61)

The eliteness-role does not distinguish between the perfectly elite and less-than-perfectly elite properties. Two things that share the property of being a dog thereby significantly resemble each other, just as two electrons resemble each other. (Though perhaps not to the same degree.) There are law-like generalizations in chemistry, biology, and psychology.[106] We can project not-perfectly-elite properties. And, most importantly, not-perfectly-elite properties in Lewis's sense will have to serve as reference magnets, if the theory is to be worth considering at all.

Take the difference between dogs and dogs-before-3000 AD. Any plausible notion of degrees of eliteness should entail that the property dogs have in common is much more elite than the property only dogs-before-3000 AD have in common. On a definitional approach to the not perfectly elite properties, this would mean that (i) the definition in perfectly elite terms for dogs is shorter than the definition for dogs-before-3000 AD, and (ii) it is shorter in the way that makes dogs *much* more natural than dogs-before-3000 AD.[107] If elite properties are reference magnets, it is easy to see why this must be the case: language users that fail to rule out properties like dogs-before-3000 AD with their use nonetheless determinately refer to dogs. Dogs should be much easier to refer to than these gerrymandered counterparts.

But there is reason to think that the definitional approach cannot deliver the goods: we shouldn't expect all of the not-perfectly elite properties properties that are reference magnets to have shorter definitions than the properties that are not reference magnets.[108] Similar points could be made by appealing to other roles connected with

[106] I will not fuss over whether these are strictly speaking called laws; regardless of the answer to this question they are generalizations that are more law-like than accidentally true generalizations.

[107] Lewis doesn't explicitly tie comparative length of definition to comparative degrees of eliteness. In principle there could be other approaches: properties that require definitions with more disjunctions, for example, are much less elite than definitions of similar length but with more conjunctions instead.

[108] Cf. Hawthorne (2006, 206), Hawthorne (2007, 434), Dunaway and McPherson (2016, 652-653).

eliteness. Should the the similarity-conferring, law-supporting, and projectable properties all have shorter definitions than the similar but intuitively gerrymandered properties that are not similarity-conferring, law-supporting, and projectable properties? This is a bold conjecture.

It is also not necessary. Lewis's approach to the perfectly elite properties is primitivist: it holds that, while mass and charge are perfectly elite, there is nothing that explains this, in a certain sense. I will develop the primitivist approach to eliteness in Chapters 4 and 5. I will also extend it to cover not only the properties that are in Lewis's inventory of the perfectly elite, but also the properties that are less-than-perfectly elite.

3.2.3 Summing up

A meta-semantic theory is a theory that gives the pre-semantic factors that make it the case that terms in a language refer what they in fact refer to. A meta-semantic theory is of special interest given the discussion of Chapters 1 and 2. If practical terms are highly stable, and refer to the same property even when used by a wide variety of possible linguistic communities, then we should aim to provide a theory of why it is that these communities, who can differ from each other in a number of ways, do not differ from each other in meta-semantically relevant respects. Of course, given the results of Chapter 2, such a theory should provide an explanation without overgeneralizing to entail **Universal Stability**.

I have given an outline of reference magnetism as a meta-semantic theory. The emphasis is on reference magnetism as a general theory of reference-determination, because any explanation of the stability of practical terms should fall out of general

principles. The theory can be summarized by the claim **Magnetism**: reference is determined by maximizing fit with use and eliteness of referents.

A community's use of a term is most naturally identified with which objects or properties they apply the term to. But there are other, potentially more important aspects to usage than individual applications. In addition, speakers will use a term in generalizations and composite expressions, and not every aspect of usage should necessarily receive equal weight in determining reference. The other component of **Magnetism** is eliteness: reference does not maximize fit with usage, if that requires assigning not-very-elite referents when other fairly elite candidate referents are nearby. Eliteness, as I have characterized it, is an objective metaphysical feature of some privileged properties. It is the feature that explains why some properties are similarity-conferring and others are not, and also feature in law-like generalizations and are projectable. In short, elite properties play the eliteness role. If **Magnetism** is correct, then it also explains why some properties are easier to refer to than others.

This is a fully general meta-semantic theory that appeals to objective metaphysical features of certain properties. So it has some characteristics that make it appealing to a realist who wishes to explain why practical terms are highly stable. But since it is a general meta-semantic theory, it should also be plausible on its own terms, and not only because of what it can do in a narrow application to practical language. In the remainder of this chapter I develop the theory in light of some objections to magnetism as a general theory of reference-determination. In Chapter 4 I turn to its application to practical language.

3.3 Objections to magnetism

The foregoing is a brief summary of reference magnetism. Where the theory has been explored directly, however, it has for the most part been in the hands of skeptics, or at least those who wish to raise difficulties for magnetism. This includes work by Williams (2007), Hawthorne (2007),and Schwarz (2014). Here I will outline some of these objections to the general project, and point to where their objections are not grounds for rejection, but rather serve to point in directions in which the theory should be amended or elaborated.

3.3.1 Definability and eliteness

Williams (2007) exploits Lewis's claim that that less-than-perfect eliteness is fixed by length of definition. Roughly, this is the idea that properties which can be given short definitions using only fundamental terms are highly elite; those that require longer definitions are to that extent not-very-elite. This has obvious bearing on a theory of reference magnetism: whatever is distant from the perfectly elite, in the sense that it requires a long definition, will to that extent be less eligible for reference.

Williams highlights the troubles that the definitional approach to less-than-perfect eliteness makes for the prospect of reference magnetism as a resource for explaining reference. There are two components to the argument. The first is a threshold claim: that there are bizarre, unintuitive candidate referents that are at least somewhat eligible. These bizarre referents pass some threshold for eligibility. The second is a possibility claim: that there are possible worlds where the intuitively correct referents are less eligible, and fall below the threshold for eligibility, while the bizarre referents remain somewhat eligible. The conclusion is that it is possible, according to a

meta-semantic theory that includes reference magnetism, that the intuitively correct referents are not the referents of a language that is very similar to ours.

This is the overridingness feature of reference magnetism, which here appears to be a liability. Williams argues for the threshold claim on the basis of the properties of a set of mathematical claims. I will not go into the details here—see Williams (2007, §3). In rough outline, the salient properties are: there is a domain of objects, which are candidate referents for terms in our language, that consists entirely of mathematical entities. And there is an interpretation of our language—which maps terms to entities in the domain of mathematical entities—that makes all of the right sentences come out true, and hence fits our usage. There is a certain degree of eliteness of the referents on this interpretation, which is determined by the complexity of construction of referents out of simple mathematical entities. Plausibly these referents are not very elite.

This is not an intuitively correct interpretation—in fact, it is much more worrisome than the claim that it is indeterminate whether our term 'dogs' refers to dogs or dogs-before-3000-AD. On the bizarre alternative interpretation Williams provides, 'dogs' refers to a construction of mathematical entities.

The eliteness of the bizarre referents is fixed from world to world: simple mathematical entities are (necessarily) fundamental, and the complexity of definitions of referents in terms of them will does not increase or decrease in various possible worlds. But Williams claims the complexity of the definition of non-bizarre referents does vary from world to world. The property of being a dog, we can assume, has a finite definition in fundamental terms in the actual world—say it is a definition of length m. There are, however, other possible worlds which are structurally identical to ours at the macro-level, including the facts about dogs and their environment. At the micro-level,

these worlds go deeper: while in our world quarks and the like are most fundamental, in other worlds the quark-like features are far from fundamental. In principle there is no limit to how much further the fundamental level can go, all the while remaining the same as ours "from the quarks up" (Williams, 2007, 338). At some point the additional complexity makes the definitions of the intuitively correct referents more complex than the definitions on the bizarre mathematical interpretation. Then we will have a world which, according to a version of reference magnetism that defines eliteness in terms of definitional complexity, is a world where a language like ours is about bizarre mathematical entities.

The arguments for both the threshold and possibility claims rest on the assumption that eliteness is tied to the length or complexity of definition in fundamental terms. The lesson is that we should reject this part of Lewis's view, not reference magnetism as a whole. Not only does it fail to fit with his overall picture, but independent reflection on the use this idea plays in the arguments for the threshold and possibility claims should make us suspect it is false.

First, the definitional approach to less-than-perfect eliteness does not fit with the rest of Lewis's view. As we have already noted, the idea that being a perfectly elite property is metaphysically primitive, while being a not-perfectly-elite property is analyzed in terms of length of definition, makes eliteness as a whole a metaphysically disjunctive kind.

Moreover, the definitional approach does not provide a notion of eliteness that plays the eliteness-role. The property of being a dog is a highly complex property from the perspective of a definition in perfectly elite terms. Doghood confers some measure of similarity, can feature in law-like generalizations, and is projectable. There are prop-

erties that have very similar definitions in perfectly elite terms; consider doghood*, which has the same definition as doghood, differing only in that it contains an extra disjunct which specifies that its instantiator is not located at a specific point in spacetime. Doghood* does not differ significantly from doghood in length of definition in perfectly elite terms. (I assume location and spacetime coordinates can be specified fairly easily using perfectly elite terms.) But doghood* plays no part of the eliteness role.

Williams's appeal to possibilities that are the same "from the quarks up" but have extra layers beneath is in general a nice illustration of the absence of a connection between length of definition and eliteness. A property like doghood in a world with extra layers beneath the quarks will play all of the components of the eliteness role just as well as doghood does in the actual world. Dogs will resemble each other in such a world to the same degree that they do in the actual world, and so on.

What this illustrates, in other words, is not that **Magnetism** is false. Instead it shows that the definitional approach to eliteness should be rejected, and reference magnetism should be implemented with an alternative conception of eliteness. There is no reason to follow Lewis's assumption that less-than-perfect eliteness is defined in terms of its length of definition in perfectly elite terms. We then should not worry that the theory will predict that intuitive interpretations are possibly less eligible than bizarre mathematical interpretations.[109]

3.3.2 More craziness

Hawthorne (2007) gives a different kind of worry about reference magnetism. He also proposes a methodological framing for reference magnetism which avoids the

[109] See Hawthorne (2006) for this idea.

worry. I will mention both, as the methodology he recommends is one I wish to take on board.

There are a number of scenarios that Hawthorne suggests create problems for the reference magnetism picture. I will focus on one, which he calls the "belief worlds" scenario:

> Suppose there is a planet where someone a lot like me is in an environment that is, qualitatively, exactly like I believe my environment to be: If I have a friend Frank whom I (rightly or wrongly) believe has just gone on a picnic in Kent, then my Twin has a friend that he calls 'Frank' who has just gone on a picnic in a place he calls 'Kent'. Consider the following "crazy" interpretation for my utterances: when I say 'Frank', I am referring to the person that my Twin calls 'Frank'; when I say 'Kent', I am referring to the place my Twin calls 'Kent'; when I say 'I', I am referring to my Twin, and so on. Predicates, by contrast, are generally interpreted in a nondeviant way: 'red' means red, 'negative charge' means negative charge, and so on. (Hawthorne, 2007, 428)

The apparent implication of the belief world scenario is related to the apparent implication of Williams's bizarre mathematical interpretation: that it is possible for reference magnetism to override other aspects of reference-determination to deliver implausible results about reference. The source of the possibility is different in this case: for Williams, whether the possibility of crazy reference obtains depends on how deep the fundamental layer of our world is; in the belief world scenario, there is another planet that corresponds to the beliefs as held by an individual speaker. While, in Williams's case, eliteness allegedly overrides reference to produce implausible results, in Hawthorne's the eliteness of the candidate referents is equal; the belief world is potentially a better fit with use, and so wins on the combined metrics in **Magnetism**. This does not have the disadvantage of relying on the length-of-definition approach to eliteness.

Of course, there is an important difference between the real world we refer to

and talk about, and the belief world (if it exists). The difference has to do with how gerrymandered our relation to the belief world is, compared to our relation to the world around us: we move around in, and interact with, the things in our world, and not the things in the belief world.[110]

Since the belief world is not causally related to us, whereas the "real" world is, only an ordinary reference scheme, and not a crazy one, satisfies such a constraint. That we should think the non-gerrymandered relation matters to a theory of reference-determination is intuitive, why it should matter is not obvious. Here is one possible answer. Reference is a relation that appears in the generalizations of a semantic theory: a typical semantic theory will involve claims of the form *speakers of language L use the expression t to refer to r*. These are law-like generalizations, and so are the kind of claims that elite properties feature in.[111] Reference, then, should be an elite property. But a semantic theory which says that we are referring to objects in the belief world makes reference out to be a gerrymandered, non-elite relation. If such a theory were correct, then speakers in possible worlds where there are no belief worlds are referring to the people and places that are right in front of them, while other possible speakers who are in identical environments refer to distant people and places, simply because they are in a possibility where a belief world does exist.[112]

[110]Hawthorne points to a limited role for causal relationships in a theory of reference-determination:

> It is clear that causal constraints can help here in filtering out certain crazy hypotheses. While it would be tendentious to suppose that there is a general causal constraint on semantic reference, it is far more plausible to think that there is a constitutive causal constraint on perception. The analogue of certain crazy interpretation problems for perceptual attention lack bite: it seems hard to take seriously the possibility that in the mirror world, the subject is perceptually attending to a mirror object. (Hawthorne, 2007, 431).

Here Hawthorne puts the point using another problem case for reference magnetism, which he calls the "mirror world". But the same points apply to the belief world, outlined above. Plausibly causation, and in particular perception, is a pretty elite relation, and so a relation to entities we are perceptually related to will, mutatis mutandis, be more elite than a relation to things we are not perceptually related to. Williams (forthcoming) provides an alternative way to build on this idea.

[111]I develop this idea more in Chapters 4 and 5.

[112]Sider (2012, 28) develops this idea in connection with the idea that semantic theories are explanatory, and

It is worth stating one consequence of this idea for **Magnetism**. If we take seriously the idea that reference should be an elite relation, there is a constraint on how we should construe the *maximization* of fit and eliteness. There are ways of "maximizing" these constraints which are very unnatural. For instance, an assignment of referents from the belief world do this. We should, instead, interpret maximization a natural, non-gerrymandered relation itself. In doing so, we rule out interpretations that, in certain contingent circumstances do well on the use and eliteness metrics individually, but do not do well on the combined score on any fairly natural account of maximization.[113]

This is not the only way to refine **Magnetism** to address these issues. In addition to adding constraints on what the notion of maximization in **Magnetism** amounts to, we could also try to be more precise about what *use* amounts to.[114] I will not catalogue all of the options here. Instead I simply wish to raise the issue of bizarre interpretations that do not rely on the definitional approach to eliteness, and highlight the role that treating reference as an elite property on its own can play in solving these issues.

3.3.3 Two projects: Lewis-interpretation and meta-semantics

A critical discussion of the merits of reference magnetism should be separated from textual considerations concerning the role of reference magnetism in Lewis's own writings. Schwarz (2014) gives a rich discussion of the relationship between Lewis's papers which inspired discussion of magnetism (Lewis, 1983, 1984), and the rest of Lewis's work on language (Lewis, 1969, 1975). Schwarz claims that, in the context in which

a semantic theory which makes bizarre claims about reference—for instance one which claims that we are referring to far-off entities and places—will not be able to explain facts about our behavior and action.

[113]Thanks to discussion from a reader here.

[114]Hawthorne (2007, §4) and Williams (forthcoming, Part II) each sketch some routes one might take in more detail.

Lewis introduces magnetism, he takes on simplifying assumings, or (in Lewis (1983)) takes on the assumptions of his opponents. The simplifying or dialectical assumptions do not appear to be a part of Lewis's own considered view. The upshot is that Lewis himself, who is often cited as the main reference point for a **Magnetism**-like thesis, ultimately rejected it.

This is interesting, and we can grant here that there is some reason to think that the thesis **Magnetism**, which resembles the picture of reference magnetism that has been inherited from Lewis in the literature, was not Lewis's own theory.[115] But here a distinction is in order. One set of issues revolves around the interpretive question, which is the relationship between **Magnetism** and the corpus of Lewis's publications on the philosophy of language. A second distinct issue is whether **Magnetism** is true. Reference magnetism can be motivated on its own terms, and these motivations can owe something to Lewis without requiring a commitment to the full package of Lewis's own views in the philosophy of language.

Since the project I am engaged in here is the second one, as it ultimately has the goal of exploring whether the realist can develop a theory of reference-determination to explain **Robust Disagreement** for practical terms, questions of Lewis-interpretation are largely irrelevant. Although the title of Schwarz (2014) is "Against Magnetism", it focuses in large part on the first issue, arguing the Lewis's own theory, for reasons sketched above, departs from a meta-semantic theory that accepts **Magnetism**. That is, as I have emphasized, an interested question; it is just not one that has much to do with whether we should be for or against reference magnetism. So it is worth moving on to consider other arguments that do bear on the overall plausibility of **Magnetism**.

[115] Williams (forthcoming) makes similar points.

Schwarz does include arguments that bear on this second project. Here is one (in this passage what he calls "natural" properties are what I have been calling "elite" properties):

> Imagine a community of language users that eat only root vegetables and a rare type of mushroom. They have a word 'food' that plays a role similar to that of 'food' in English, which they apply to root vegetables as well as the mushroom. But the root vegetables by themselves form a much more natural category than the root vegetables together with the mushroom. According to reference magnetism, the community's word 'food' might therefore pick out only root vegetables, since that is the more natural referent. How does this help explain or systematize the community's linguistic practice? What is the explanatory advantage of the magnetic interpretation? (Schwarz, 2014, 31-32)

The objection is that magnetism will objectionably entail an implausible referent for the term 'food' in this community. There are several ground-clearing points to make before turning to an evaluation of this argument.

First, granting that root vegetables are elite, but root-vegetables-and-mushrooms is not, **Magnetism** is not thereby constrained to assign root vegetables as the referent of 'food'. The theory says that it there are two components to reference-determination. Eliteness is one, but use also matters. Moreover use extends beyond the simple application of terms to particular items. The generalizations in which a community uses their terms also matter. And a community that uses 'food' with the food-role will accept statements like 'food is nutritious' and 'restaurants serve food'. Since mushrooms are nutritious and are served at restaurants, it is not obvious that reference magnetism will override use in this case. It is not a consequence of **Magnetism** that we *always* refer to elite properties; eliteness is one of two reference-determining factors.

Second, the actual use of the community is not the only relevant issue. We should also look at how the community is disposed to use the term: if, for example, they

discovered citrus fruit and found it edible and tasty, would they apply the term 'food' to it? The suggestion that they use 'food' with the food-role suggests that they would. This disposition (and others like it) are clearly relevant to what the community is talking about with their term 'food'. Just as a focus on individual applications of the term and not the generalizations it features in results in an impoverished conception of use, so does a focus on actual usage patterns at the expense of the dispositions to use the term.

Finally, we should reject that a theory of reference magnetism is forced to choose between root vegetables and root-vegetables-and-mushrooms as the only two candidate referents. Given this choice, it seems plausible that the magnetism component of the theory will favor root vegetables as the more elite candidate. But these aren't the only candidates. Taking the role and dispositions to use 'food' seriously, there are other candidates as well. And some of them are pretty elite: nutritious organic material, for example, would fare quite well on eliteness considerations.

What should we say about Schwarz's example? The right answer depends on the details of the imaginary community. If they do use the term with the usual role and dispositions (i.e., if they only actually eat roots and mushrooms, but accept generalizations about food and have dispositions to use it like English speakers in which they eat other substances) then they will end up talking about what we talk about with 'food'. Another way to fill out the case involves the community insisting—both in actual practice and in dispositions in nearby scenarios—that only root vegetables and certain mushrooms fit the label 'food'. This community rejects, or is disposed to reject, generalizations such as 'anything nutritious is food'. Then we should say that their usage overrides the eliteness of other candidates.

But even in this second case we should not hold that eliteness has nothing to do with reference. It is an important fact about the case that we are presented only with root vegetables and root-vegetables-plus-mushrooms as the only candidate referents. One thing that is certain is that the community does not use 'food' to talk about gerrymandered referents such as the root vegetables outside the Bermuda Triangle or the root vegetables except time-slices at exactly 12:01 and 39 seconds on January 2nd. We can get as specific as we like about the conventions and practice of the imaginary community, but there will be some limits to the conventions, and we can always find a gerrymander that deviates where the conventions leave off. By framing the theoretical choice as one between root vegetables and root-vegetables-plus-mushrooms, we have already made use of the explanatory advantage afforded by reference magnetism, since the framing already assumes that some vastly more gerrymandered properties are not eligible.

Here is a second objection from Schwarz. Eliteness, according to reference magnetism, can sometimes override use, and one possible example is the following:

> Perhaps our ancestors used 'fish' with the convention that it is to pick out a biologically homogeneous class of objects. We can imagine that the word was introduced by a stipulation to the effect that it picks out the biologically most natural kind including such-and-such exemplars (here we point at some carp and herrings) and excluding such-and-such others (crabs, snails, elephants). Since the fish form a much more natural biological kind than the fish together with the whales, this would mean that whales do not fall under the predicate 'fish', irrespective of whether anyone is aware of that fact. (Schwarz, 2014, 32)

Schwarz notes that this seems to provide a motivating case for reference magnetism. On a suitable filling in of the details, this seems plausible. But then he argues that the motivation is illusory:

> The reason why naturalness here plays a role in determining reference is that *this is how we use the words*. Nothing forces us to use words like 'fish' and 'temperature' with the convention that they pick out reasonably natural properties. Moreover, there is nothing special about naturalness here. All kinds of features can enter into the truth conditions associated with sentences, and thereby (derivatively) into the application conditions of predicates. Our linguistic practice makes 'fish' pick out a biologically natural class of things, but also a class of things whose members typically have fins and live in the water. For other terms, objective naturalness is irrelevant. (Schwarz, 2014, 33, his italics)

The claim here is that, when eliteness does play a role in determining reference, it does so because speakers intend it to do so. It operates like any other feature of use. If it is true, then **Magnetism** is false, since it says that eliteness is an additional component of reference-determination, and operates regardless of whether speakers have the intention to refer to elite properties.

This point assumes that speakers already have a term 'natural' (or, in present jargon, 'elite') that refers to eliteness, or at least a shared concept of the property, and can thereby include the term/concept as a part of the reference-determining use-profile of their terms. For instance, speakers can say (or think) things like 'the term 'fish' refers to the most elite biological kind shared by trout, salmon, and cod'. Perhaps it is plausible that actual speakers have a term which allows them to formulate these claims, and thereby to constrain the eligible referents for their terms to elite properties. But **Magnetism** should be understood as a *necessary* claim about how reference is determined. It applies to possible linguistic communities, including those that do not have a term or concept that refers to elite properties. In order to assess the plausibility of the claim that eliteness only constrains reference as an aspect of the usage of the term 'elite', we need to ask what these possible communities who lack such a term are referring to.[116]

[116]It is also worth reflecting on the fact that, when communities do refer to eliteness, they refer to eliteness, which is presumably itself elite, rather than one of the many nearby eliteness-like but gerrymandered properties. Cf. Dorr and Hawthorne (2013).

Consider a community of very simple creatures who live in a very simple perceptual environment. They encounter only square objects, round objects, and triangular objects. Their simple cognitive lives do not enable them to have very rich descriptions of what is going on. But they can demonstrate via perceptual attention the shapes in front of them (expressed with 'there is') and they can use one of three predicates: *S*, *C*, and *T*. When squares appear, they routinely say 'there is a *S*', when circles appear they say 'there is a *C*', and when triangles appear they say 'there is a *T*'. They don't apply the predicates to other shapes (*S* is never uttered in front of triangles) and have no other cognitive resources. Importantly, they can't say or think things like 'all *S*s have a very elite property in common'.

If this pattern goes on for a couple of years, we would obviously say that *S* refers to squares, *C* refers to circles, and *T* refers to triangles. Ex hypothesi they have no concept of eliteness to ensure that they are interpreted as referring to squares with *S* rather than some property that coincides with the squares for the few years of use and then deviates to include triangles in the future. And if they sometimes slip up and apply *S* to a triangle here and there, we still would interpret them as referring squares even though they don't possess a concept of eliteness that allows them to say that *S* refers to the most elite property in the vicinity of their use.

The motivations for magnetism are present in full force, even if some communities are in a position to use the term 'elite' as part of their use of other terms in their language. Since **Magnetism** should be understood as a necessary truth about reference-determination, motivations for the thesis are not defeated even if we grant that actual speakers manage to restrict candidate referents to elite properties simply in virtue of their use. When this occurs, reference to elite properties is simply overdetermined.

This chapter has set the stage to return to a realist-friendly explanation of the disagreements between possible users of practical language. I have defended **Magnetism** as a general theory of reference-determination, which applies to both descriptive and practical terms. This is because **Magnetism** says that both usage of a term and the eliteness of candidate referents play a role in determining what a term refers to. Since both descriptive and practical language have distinctive patterns of use, and candidate referents for both descriptive and practical terms can be elite or not, **Magnetism** in principle can provide an explanation of why both descriptive and practical terms have the referents they in fact have.

Applying the theory in a plausible way requires a more sophisticated understanding of both use and eliteness. Use includes not only the individual applications of terms in a language, but also the generalizations that speakers accept, and the compositional structure of the expressions that they appear in. Eliteness is a metaphysical category, which (according to **Magnetism**) plays a role in reference-determination. But it also does other jobs as well: it explains similarity, distinguishes genuinely law-like generalizations, and confers projectability on certain properties. As I argue in the next chapter, this is the kind of feature a realist should claim is instantiated by moral and normative properties. Defending **Magnetism** requires some further developments. Eliteness should not be tied to definitional length; *maximization* should make reference an elite and not gerrymandered relation, and the constraints on reference it puts forward should be understood as necessary and not contingent.

Given the assumption that moral and normative properties are elite, the realist can use **Magnetism** as I have developed it to explain why **Robust Disagreement** is true. Moreover the theory very naturally fails to generalize to entail

Universal Disagreement. Thus not only is reference magnetism a plausible general meta-semantic theory, and one that it is natural for a realist to adopt, it explains the facts about the extent of possible disagreements with practical terms, while also explaining why the appropriate limits are in place. I turn to developing these claims in the next chapter.

Chapter 4.

Magnetism and practical terms

Chapter 3 defended and developed reference magnetism as a simple, general theory of what determines reference for a language. It relies on two claims. The first is that some properties are metaphysically elite, as they are objectively distinguished from non-elite, gerrymandered properties. The second is that the elite properties are easy to refer to: reference is determined by what maximizes fit with use and eliteness.

On the face of it, magnetism is a candidate to explain the stability of practical terms. If properties like obligation and moral rightness are highly elite, then they are highly eligible referents for moral and practical terms. It is in principle possible for two communities to use their normative term 'ought' very differently, and yet be such that the highly elite property of obligation best satisfies the constraints on reference for their respective languages. They are talking about the same property, and are capable of having a substantive disagreement.

The possibility appealing to reference magnetism to explain Moral Twin Earth-style disagreement has not gone unnoticed.[117] But the proposal has been criticized on a number of points.[118] While there are plausible responses to these criticisms, these have not been developed at length. Moreover the existing debate surrounding reference magnetism for practical terms cannot be separated from what reference magnetism should try to explain. Many of the main discussants emphasize the relative stability of practical terms.[119] But the plausible limits to this stability, which make

[117] See van Roojen (2006), Edwards (2013), and Dunaway and McPherson (2016).
[118] Schroeter and Schroeter (2013), Eklund (2017), Williams (2018)
[119] Eklund (2017), Williams (2018, forthcoming)

Universal Stability—and by extension **Universal Disagreement**—false do not receive equal emphasis.

This has two related, but ultimately distinct, consequences for any discussion of the relationship between reference magnetism and realism about morality and normativity. The first is that a meta-semantic proposal by the realist that includes reference magnetism should not be rejected simply because it does not entail **Universal Disagreement**. Since there are some possible communities who use practical terms with the same characteristic role, but fail to have substantive disagreements with each other, it is not a strike against reference magnetism that it entails that some communities are referring to different properties. Of course much hangs on the details: we need an inventory of which specific possible communities are having substantive disagreements. If reference magnetism fails to explain why these communities are talking about the same thing, then it is a problem for the theory. But a failure to entail **Universal Disagreement**, on its own, is not.

There is a second consequence of the failure of **Universal Disagreement** for any discussion of reference magnetism. It is a prima facie explanatory desideratum for any meta-semantic theory—reference magnetism included—that it should explain *why* **Universal Disagreement** fails. Explaining the cases where possible communities do not disagree with each other is just as important for a meta-semantic theory as explaining the cases where they do disagree.

This chapter delivers part of the explanation. §1 develops the connection between realism about morality and normativity, and the eliteness of moral rightness, obligation, and other properties. §2 responds to objections to reference magnetism as a meta-semantic theory for practical terms. §3 applies **Magnetism**, coupled with as-

sumptions about which properties are elite, to explain why practical disagreement is robust, but not not universal. §4 extends this picture to the contextualist semantics.

4.1 Realism, eliteness, and primitivism

What does the notion of metaphysical eliteness—and in particular the idea that properties like moral rightness might be metaphysically elite—have to do with realism? How should the realist conceive of the metaphysics of elite moral and normative properties? In other words, what (if anything) makes these properties elite? And how should the realist conceive of the relationship between the eliteness of these properties and the reference of practical terms, if **Magnetism** is true? The last chapter defended **Magnetism** as a general theory; here I begin by motivating the view that, for a realist, moral and normative properties have the metaphysical status to serve as reference magnets.

4.1.1 Eliteness and realism

The notion of eliteness—an objective metaphysical feature that distinguishes some privileged, explanatory properties from other non-explanatory, gerrymandered properties—should be a welcome resource for realists about morality and normativity. The view that a realist about morality and normativity holds that moral and normative properties are metaphysically privileged in some sense is a common one in the literature. For example: Fine (2001) holds that ethical properties are a part of Reality (see also Wedgwood (2007)), McPherson (2015) argues that for the realist, normative properties carve at nature's joints, and Leary (2017) holds that normative properties have essences that explain their relationship to natural properties.[120] I will not defend

[120]See also Dunaway (2017c, MS) for related ideas.

realism here. It is, however, worth saying something about why the commitment to the eliteness of properties like moral rightness and all-things-considered obligation is a natural commitment for the realist to have, before turning to specific applications of eliteness to a theory of reference for practical terms.

The slogan at the beginning of this book that captures realism is Matti Eklund's: "reality favors certain ways of acting".[121] One function of slogans like this is to distinguish realism from non-cognitivist and subjectivist views. On some non-cognitivist views, practical language doesn't primarily have the function to represent something outside of speakers and agents at all: rather it expresses or conveys their attitudes about how to act.[122] A subjectivist holds that practical language does have a representational function, but that what it represents is simply an agent's own attitudes or desires.[123] On either kind of view, the rightness of an act is constituted or explained by what we think about it, or how we value it. This is not realism about morality.

A realist instead holds that it is a feature of an act itself that makes it morally right, or obligatory. Giving to the poor is not obligatory because of how we think of it; we would be obliged to give to the poor even if we had no inclination to do so. This is the mind-independence of obligation; often mind-independence is assumed to be the characteristic feature of realism.[124] But realism requires something more: that rightness and obligation are not gerrymandered and "unnatural" in Lewis's sense. It is a mind-independent fact that Cicero held a consulship in a year number not divisible by 17, but there are similar and equally gerrymandered properties that Cicero does not

[121] Eklund (2017, 1).

[122] Gibbard (2003), Blackburn (1984) are the classic articulations of this view; Dunaway (2016) argues for its anti-realist implications. The relationship between expressivism and realism is complicated; I will not try to provide a full account here. See also Dunaway (2010).

[123] See Schroeder (2010) on the details of subjectivism; Dunaway (MS) sketches why this view counts as anti-realist.

[124] Dunaway (2017c)

instantiate. The realist should think that the obligatoriness of giving to the poor does not involve an arbitrary, gerrymandered property.

This is the connection between eliteness and realism. If the obligatoriness is elite, then obligation has a robust explanatory role that it makes sense to describe as a feature of reality. Recall the roles that eliteness plays: the eliteness of a property explains why it confers similarity on its bearers, why it features in law-like generalizations, and why it is projectable. These are not only objective, mind-independent features of a property; in addition they are also features that are not shared by the quite plentiful arbitrary and gerrymandered properties.

The notion of eliteness thus provides the realist with one way of articulating the idea that there is some genuine, non-arbitrary requirements on how to act, and that these requirements are features "in reality" and not products of how we think about our obligations. I am not going to argue here that this is the only way of articulating the realist view, but it is promising, and it is the version of realism I will be developing here. The connections with reference magnetism will already be clear at least in outline, and so it is an especially promising version of realism to pursue in connection with facts about the semantic stability of practical terms.

The eliteness of moral and normative properties is related to another metaphysical issue that has interested realists. This is the individuation of properties, and especially the relationship between moral and normative properties, and their supervenience bases.[125] There are various possible views about how properties should be individuated. On one conception, they are individuated *intensionally*, and so any properties that are instantiated by all and only the same individuals at every possible world are

[125]I thank an anonymous reader for raising this question, and for noting its relevance to several issues that arise below, connected to reference magnetism.

identical.[126] Others hold that properties are individuated *hyperintensionally*; that is, they hold that some properties which share an intension are distinct.[127]

The hyperintensional view is, strictly speaking, simply a view about the distinctness of some properties that share an intension, but in application it usually involves the additional claim that, among the hyperintensionally individuated properties are moral rightness its supervenience base. For instance, supposing rightness supervenes on the property of maximizing happiness, the hyperintensionalist will hold that rightness and happiness-maximization are distinct, although necessarily they are instantiated by the same acts. What grounds we have for individuating properties in this way is not obvious, but some account here is needed if individuation is a significant metaphysical issue.[128]

I will not pursue this question here, because a metaphysics of eliteness for moral and normative properties can be implemented in a framework where properties are individuated intensionally, or in a framework where they are individuated hyperintensionally. Some adjustments will need to be made depending on the choice: a hyperintensionalist can claim that rightness is elite while happiness maximization is not; an intensionalist must hold that if rightness is elite, and is identical to happiness-maximization, then happiness-maximization is elite as well.[129] In some places the choice of framework will raise special problems: I will note them when they arise. But, in general, the point for present purposes is that the realist can pursue a view on

[126]This is the view in Jackson (1998).
[127]See Shafer-Landau (2003). For more discussion see Streumer (2008), Suikkanen (2010), and Dunaway (2015).
[128]Bader (2017) provides one. The criterion cannot be mere difference in cognitive significance: 'right' and 'maximizes happiness' convey different information, but so do many necessarily equivalent descriptions, including 'triangle' and 'three-sided closed plane figure'. Property non-identity cannot carry much metaphysical significance if properties are so easy to come by that even triangualrity and being a three-sided closed plane figure are distinct.
[129]Dunaway (2015) outlines the latter view in more detail.

which rightness and obligation are metaphysically elite properties, and leave for further investigation whether these properties are identical to their supervenience bases.

4.1.2 Primitivism and arbitrariness

Chapter 3 outlined the eliteness-role: elite properties confer similarity, are the constituents of genuine law-like generalizations, and are projectable. The eliteness-role provides an objectively significant role that distinguishes a property from mind-dependent and uninteresting properties. A realist will be interested in an account that puts moral and normative properties in the former category.

That rightness has a property which plays the eliteness-role is a substantial metaphysical claim, and is not trivial. But even if we accept it, there is a further metaphysical question that can be raised: what makes it the case that rightness is elite, which is to say, what makes it the case that rightness has a property which plays the eliteness-role? (Analogous questions can be raised for any other allegedly elite property.) It might seem that we need to answer this question. After all, eliteness plays a significant explanatory role, as it (among other things) makes some properties reference magnets. This raises to salience the job of saying which properties are elite, in order to derive conclusions about which properties we, or other possible linguistic communities are referring to. While it might be tempting to claim on this basis that we need an account of what makes a property elite, I will argue that the temptation can be resisted.

Return to the bifurcated view of eliteness from David Lewis: on his view, the notions of *perfect* eliteness and *less-than-perfect* (or *degreed*) eliteness look very different from each other, metaphysically speaking. Lewis took the fundamental quantities from

physics such as charge and mass to be good candidates for perfectly elite properties.[130] For Lewis, there is nothing that makes these properties perfectly elite. Properties that are elite-to-some-degree, on the other hand, are so in virtue of their relationship to the perfectly elite. This is Lewis's definitional approach: the property of being a planet is, for instance, less elite than the property of being a hydrogen atom since the former has a very long definition in perfectly elite terms. Perfect eliteness is, on this view, *primitive* in the sense that it lacks further grounds. Degrees of eliteness are not.

We don't, however, need to follow Lewis by treating the elite properties from physics differently from other instantiators of eliteness. Chapter 3 covered one compelling reason to do so; taking degrees of eliteness to be grounded in length of definition will produce implausible results when combined with **Magnetism**, as Williams (2007) shows. Rather, a unified approach is possible. This is the primitivist approach to eliteness, which takes all facts about eliteness, and not just the perfectly elite properties from physics, to be primitive.

If moral rightness and all-things-considered obligation are elite, then on the primitivist approach, it is an ungrounded fact that these properties are elite. This opens up the possibility of using **Magnetism** to explain meta-semantic facts about practical terms, without having first to do any special theorizing about the nature of rightness and obligation, in order to first show that they are, in fact, elite. This is the approach I will pursue, but it will help to first clarify the primitivist approach by considering an immediate objection. Obviously, not every property is elite: the point of a theory of eliteness is to distinguish between properties which are metaphysically privileged and those which are not. But on the primitivist approach we face a worry that any

[130] Lewis (1983, 356-7)

theory of which properties are elite will be arbitrary and unmotivated. Take the property of being a planet, which is presumably fairly elite. A theory which claims that planethood is more elite than nearby gerrymanders cannot claim that this is because only planethood instantiates the grounds of eliteness. Ex hypothesi there is nothing that grounds the eliteness of planethood on the primitivist approach. So what does motivate claims about eliteness, or are all such claims arbitrary?

There is already a solution to this worry in Lewis's original theory of the perfect eliteness of physical properties like charge and mass. The eliteness of these properties is not grounded in anything. This is an especially important point for an application of reference magnetism to practical terms, since the definitional approach looks unpromising as a way of making moral and normative properties come out as relatively elite. It is not, however, ad hoc or arbitrary, on Lewis's view, to hold that charge is elite and its nearby gerrymanders are not. Charge plays a role that the other gerrymanders do not. This is the eliteness-role.

That charge plays the eliteness-role is an objective fact, and claiming that it does so is not to arbitrarily distinguish between charge and other properties. It is clear, for instance, that two negatively charged electrons resemble each other to some degree; the same cannot be said for two things that share the gerrymandered property of either being negatively charged or located in California. Similarly for other components of the eliteness-role: negative charge shows up in the law-like generalizations of physics, and is projectable.

That a particular property is elite entails that it has a property which plays the eliteness-role. We should not treat the identification of a role as another attempt at grounding eliteness. Eliteness is supposed to be what part of what explains why a

property is similarity-conferring, or why it features in law-like generalizations. That is: two electrons are similar in part *because* they share the elite property of being charged. Similarly, that negatively charged objects repel each other is a law in part *because* charge is elite. Eliteness is part of what explains why charge plays these roles. These roles cannot in turn provide a definition of eliteness—this would be to get the explanatory order backwards.[131]

That elite properties can be identified on a principled basis because they have a property that plays the eliteness-role is compatible with primitivism about eliteness. The explanatory order is: first, such-and-such property is elite, which is a primitive, ungrounded fact; second, because such-and-such property is elite, it confers similarity, features in law-like generalizations, and is projectable—and so plays the eliteness-role. The first sustained attempt to exploit reference magnetism in a meta-semantics for realists about morality can be found in van Roojen (2006). There, van Roojen appealed to a notion of "discipline-relative eliteness", acknowledging distinct properties of *being physically elite*, *being biologically elite* and, importantly, *being morally elite*.[132] Properties that instantiate any of these discipline-relative eliteness properties serve as reference magnets, on this view. Primitivism as I have developed it departs from van Roojen's picture, first, by recognizing only one notion of eliteness, which possibly comes in degrees, and which is instantiated by physical, biological, and moral properties (among others). Second, discipline-relative eliteness is plausibly best understood (though van Roojen does not say this explicitly) as non-primitive: since the properties that appear in the laws of different sciences all instantiate different kinds of eliteness, it is natural to hold that something grounds, or explains why, a given property instantiates one

[131]Cf. Dunaway (2016)
[132]The terminology is mine. Van Roojen uses the term 'discipline-relative naturalness'.

kind of eliteness and not others.[133] The primitivist view that I have outlined officially departs from van Roojen's picture on both fronts. But it follows van Roojen in holding that there is a connection between eligibility for reference and appearance in the law-like generalizations of a discipline.

Before developing a meta-semantics for practical terms that includes both reference magnetism and primitivism about eliteness, I will sketch some options for how the theory can handle what Lewis called the not-perfectly-elite properties. There are multiple options here, which I will not choose between, but only sketch the costs and benefits.

4.1.3 Primitivism and reference magnetism

I am proposing that we treat all facts about eliteness, including the realist's elite of rightness and obligation, as primitive. Lewis's proposed definition of degrees of eliteness is incompatible with the roles eliteness is supposed to play. But this does not mean that there are no benefits to the definitional approach, and which are benefits we forfeit as primitivists.

Lewis notes that the thesis that there are degrees of eliteness (or 'naturalness' in Lewis's jargon) is theoretically fruitful for reference magnetism:

> There is the line between the perfectly natural properties and all the rest, but surely we have predicates for much-less-than-perfectly natural properties. There is the line between properties that are and that are not finitely analysable in terms of perfectly natural properties, but that lets in enough highly unnatural properties that it threatens not to solve our problem. We need gradations; and we need some give and take between the eligibility of referents and the other factors that make for 'intendedness' [...] Grueness is not an absolutely ineligible referent (as witness my reference to it just now) but an interpretation that assigns it is to that extent inferior to one

[133] A natural candidate for the explanans is the discipline-specific laws: the fact that mass appears in the laws of physics grounds the fact that mass is physically elite, and so on.

> that assigns blueness instead. Ceteris paribus, the latter is the 'intended' one, just because it does better on eligibility. (Lewis, 1983, 372)

Here the point is that there are properties which are not perfectly elite—blueness is an example—which must be more eligible for reference than other not-perfectly-elite properties, such as grueness.[134] This provides one rationale for having a distinction between degreed and perfect eliteness: since properties can be eligible for reference to varying degrees, there are many properties that fall under the heading of 'degreed eliteness'. Perhaps, for each real number n, there is the property *elite-to-degree-n*.

The primitivist does not acknowledge that degrees of eliteness are determined by any definitional facts. But there are various approaches the primitivist might take to degrees of eliteness, which balance of the benefits degrees of eliteness yield for reference magnetism, with the costs of theoretical complexity. Some of these costs are metaphysical: they saddle the primitivist approach with many (possibly infinitely many) primitive degreed eliteness properties. Others are the consequence of complications of trying to make do with a parsimonious metaphysics of eliteness, which complicates theorizing elsewhere.

Broadly, the options fall into three categories. All of these options will be available to a version of primitivism which holds that moral and normative properties are among the elite, or elite-to-some-degree, properties. I will not attempt to decide between them, but rather simply note that the theory I develop here will need to take on some costs, in exchange for the benefits of rejecting Lewis's definitional approach.

1. There are no degrees of eliteness. The simplest option is to hold that there is no difference between the eliteness of the fundamental physical properties like mass and

[134]Following Goodman (1955), an object is grue iff it is green and examined before the year 2000 or blue thereafter.

charge, and the eliteness of higher-level properties like hydrogen and organisms. All of these properties are elite simpliciter.[135]

One cost to this approach is that a meta-semantic theory that incorporates reference magnetism will have fewer resources to work with. As formulated by **Magnetism**, reference magnetism is the idea that reference is determined by maximizing two things: fit with use, and eliteness of the assigned referent. On this approach there are only two kinds of candidate referent: those that are elite (simpliciter), and those that are not elite at all. At some point, eltieness gives out. For instance we need to be able to explain why it is that we refer to grueness, even though it is presumably not among the elite properties. This is Lewis's motivation for degreed eliteness: even if grueness is not very elite, we still do manage to refer to it, and so we should grant that it possess some small degree of eliteness, and so is easier to refer to than other properties that are even more hopelessly gerrymandered.

This worry is not insurmountable. 'Grue' is defined in other terms. If these terms stand for elite properties, or can are used with connections to other terms that stand for elite properties, then a simple on-or-off eliteness view might explain why 'grue' refers to grueness and not other nearby properties, even though none are elite.[136] Of course 'grue' is a special case, as it is introduced with a stipulated definition. Extending this strategy to explain cases of reference to non-elite properties with terms that do not

[135]There is an analogous view in Schaffer (2004), though in this paper Schaffer is not interested in elite properties, but rather in the existence of *sparse* properties. This is a different framework: whereas in the present framework we hold that charge is an elite property, while the gerrymandered charge* is not, Schaffer would say that charge exists, while charge* does not. Schaffer considers supplementing the list of sparse properties with special science properties like being an organism. Just as existence is an on-or-off notion, so is eliteness on the present proposal. Structurally they will have many of the same features.

[136]Here is one simple model on which this picture might work. Suppose that we use 'green' and 'blue' by saying things like 'green things cause green-sensations in humans' and 'blue things cause blue-sensations in humans'. If 'cause', 'green-sensations', 'blue-sensations', and 'humans' all refer to elite properties, then use of 'green' and 'blue' will refer to properties that cause green- and blue-sensations in humans. Grueness, although not eligible at all, is a property that is cauJamie related to other properties that are eligible, and this is what allows us to talk about it.

have stipulated definitions may prove difficult. A theory of reference magnetism that discards degrees of eliteness in favor of a generous distribution of eliteness simpliciter will have to face this issue: even if many properties are elite, there will be cases where there are no elite candidate referents available. Such a theory may have to tolerate a significant amount of referential indeterminacy in these cases.

2. Non-perfect eliteness exists, but these properties are purely relational. Let a *purely relational* fact be a fact that is most fundamentally expressed by a relational term like 'is more than' or 'is less than'. This is to be contrasted with a *degreed* fact, which can be most fundamentally expressed with reference to numerical values. For example: the relationship between the height of two people is not a purely relational fact; it is degreed. That Jamie is taller than Sue is a relational fact. But it is not purely relational: the relationship between their heights depends on the fact that Jamie is m inches tall and Sue is n inches tall, where $m > n$. The fact that Jamie is taller than Sue depends on, and is less fundamental than, this fact about the relationship between their numerically quantified heights.

Not every relational fact depends on a degreed fact in this way. A tournament-style competition in which a winner and runners-up are determined by a series of head-to-head matchups determines which competitor finishes first, second, and third. The winner stands in the *finishing ahead of* relation to all other competitors. But there is no sense in asking by how much the first-place finisher finished ahead of the second- or third-place finishers in a competition with this format. The rankings are purely relational and do not depend on any degreed facts.[137]

The eliteness of non-physical facts might be purely relational. That is: charge is

[137]By contrast the finishers in a footrace have degreed properties such as *finished the race in n seconds* which determine the relational facts about who finished first.

perfectly elite; hydrogen is less elite, and organisms are even less elite than hydrogen. At the other end of the spectrum we have properties like being grue, being a dog-not-owned-by-Cicero, and even more gerrymandered and less elite properties. Obligation on the realist view is somewhere on the scale of eliteness—where, exactly is a good question.

Crucially on this picture there is no precise degree to which hydrogen, organisms, or the property of being grue are elite. All we can say about them is which properties they are more elite than, and which properties they are less elite than. A primitivist will treat these relational facts about eliteness as basic and ungrounded: there is nothing in virtue of which charge is more elite than hydrogen. These are the primitive eliteness facts, when we wed the primitivist approach to the view that eliteness is purely relational.

This view has something to say directly about why grueness is not very eligible, but more eligible than other referents in the area. While it is highly gerrymandered, there are other, even more gerrymandered properties: for instance, the property of being grue-and-not-located-on-Saturn. This property is even less elite. So it is even more ineligible as a referent.

Purely relational eliteness-facts come with costs as well. The first is that commitment to a range of primitive purely relational eliteness facts is a significant metaphysical commitment. Primitive eliteness (simpliciter) is a commitment that can be justified by sufficiently robust explanatory power. If the eliteness facts are purely relational, then they are more complicated: they are instantiators (or have relata) that go beyond the simply elite.

In addition this extra metaphysical commitment does not make reference mag-

netism a simple theory. If the eliteness facts are purely relational, there is no answer to the question of *how much* more elite hydrogen is over its gerrymandered neighbors. We can only say which things hydrogen is more elite than. When it comes to reference-determination, hydrogen will count as the referent of a term only if it maximizes fit with use and eliteness. We know that hydrogen should win out over candidate referents that are less elite that fit use equally well. But magnetism is also supposed to play an *overriding* role in some cases as well: a highly elite referent can fit use less well than other candidates, but still be what we are talking about, owing to eliteness. How this overriding happens is somewhat mysterious if there is no fact about the degree to which the highly elite referent beats out its competitors on this front.[138]

3. Non-perfect eliteness exists, and these properties are degreed. A final option is to hold that the facts about non-perfect eliteness are degreed. That is, it is not only a fact that hydrogen is more elite than organisms; in addition there are facts about precisely *how* elite each property is. When Jamie is taller than Sue, there is in addition a fact about how tall Jamie is, i.e., the fact that Jamie is m inches tall. Similarly, according to this view, hydrogen is not only more elite than organisms. There are specific degrees n and m such that hydrogen is elite to degree n, organisms are elite to degree m, and $n > m$.

A primitivist approach to eliteness which takes the eliteness facts to be degreed is metaphysically more costly. For the primitivist, there is nothing that grounds the eliteness facts. If eliteness is degreed, then there is not just one kind of fact that is ungrounded: rather, for every degree of eliteness there is a distinct primitive fact about which properties are elite to that degree. On the first two implementations of

[138]Of course we have already argued in Chapter 3 that *maximization* in **Magnetism** should be a very elite relation, since reference is not gerrymandered. Perhaps it could be argued that the relational eliteness view does not make reference magnetism significantly more mysterious than it is already.

the primitivist view, there is just one kind of property that is ungrounded: eliteness (simpliciter), or the relational property of *being more elite than*. But with primitive degreed eliteness, things are more complicated. There are primitive facts about which things are elite-to-degree-*n*. Likewise, since *m* is a distinct degree, facts about what is elite-to-degree-*m* are another class of primitive fact, and similarly for every other degree.

The metaphysical costs come with a meta-semantic benefit: if eliteness is degreed, then the workings of reference magnetism will be straightforward. If reference is determined by maximization of fit with use and eliteness, both elements of reference-determination will have the structure to yield determinate predictions about reference. Presumably the degree to which a candidate referent fits with use of a term can be quantified. If eliteness is degreed, then there is a fact about the degree to which each candidate referent is elite. The referent of that term, according to **Magnetism**, is simply the candidate referent that maximizes these two values.[139]

4.2 Objections to reference magnets for practical terms

Primitivism is arguably a necessary component of any view that wishes to apply reference magnetism to the meta-semantics of practical terms. It can be implemented in a number of ways, which I sketched above. I now turn to defending its application to practical terms from objections.

In various places in the literature it has been objected that reference magnetism is a non-starter as a meta-semantic theory for practical terms. Here I will argue that these

[139]This picture simplifies in a number of respects. As we have noted previously, a serious theory of reference-determination will not settle the reference of *t* on its own by maximizing fit and eliteness; it will determine reference by maximizing across the entire language of which *t* is a part. Also the maximization function itself is left opaque here. Perhaps simply adding up the degrees of fit and eliteness does not determine the interpretation that maximizes these twin considerations, because the result will be a gerrymandered reference-relation—Cf. Chapter 3.

objections are mistaken, because they fail to consider the best version of reference magnetism for practical terms. For simplicity, I will conduct the discussion using eliteness simplicter. Officially, however, my replies, and subsequent positive development of the view in the next section, are neutral between the various primitivist approaches to degreed eliteness.

4.2.1 Magnetism and overriding definitions

In Chapters 1 and 2 we qualified the **Universal Disagreement** thesis in a number of ways. One important consideration is that **Universal Disagreement**, even for those who accept it, is should be understood as a thesis that covers only possible uses of *simple* practical expressions. Some possible communities use practical terms but also treat them as equivalent to other complex expressions. **Universal Disagreement** is a non-starter if it is understood to entail that practical roles overrides definitional associations, so that even communities that use defined practical expressions are capable of disagreeing with other possible users of practical language.

Here is a concrete example of a community that uses a practical term as a defined, and not a simple, term. Consider a community that speaks a language including the terms 'maximizes' and 'happiness', which mean the same things as in English. The community might form a complex term out of these expressions, 'maximizes happiness'. Compositional rules for the language are the same as those for English, so this expression refers to the property of maximizing happiness. So far this community is identical to an English-speaking community. The important difference, we can imagine, is that this community *also* uses the term with a practical role. For instance, suppose they associate the Gibbard role with the term: they treat someone who ap-

plies 'maximizes happiness' to one of their options, and yet fails to perform the act, as making a conceptual mistake. There is nothing incoherent about this use. They might even introduce a simple term, 'ought$_U$', which they treat as conceptually equivalent to 'maximizes happiness', and so has both the same reference and normative role. This community is using their term 'ought$_U$' to refer to the property of happiness-maximization; after all, the community treats 'ought$_U$' as definitionally equivalent to 'maximizes happiness'.

Some have raised a worry that reference magnetism will predict the wrong result about this case. If we consider a community that uses a simple normative term, for which the composite term 'maximizes happiness' plays no role in determining reference, then reference magnetism predicts that if happiness-maximization is not elite, the community is not referring to happiness-maximization. They are instead referring to the elite normative property, whatever that turns out to be. We might then worry that the same goes for a community that uses the defined normative term 'ought$_U$'.

Schroeter and Schroeter (2013) make a claim along these lines, claiming that the possibility that reference magnetism overrides use will yield an implausible theory of reference for practical terms. They say:

> [M]aking perfectly natural properties into sufficiently strong reference magnets threatens to make it impossible to refer to any less-than-perfectly-natural properties in the vicinity of moral rightness. Say you introduce names for the properties picked out by two non-coextensive moral theories, one Kantian, one utilitarian. A strong naturalness constraint may lock both predicates ('is right$_K$' and 'is right$_U$') onto the very same perfectly natural property rather than two distinct less-than-perfectly-natural properties. The problem here is the inverse of that faced by descriptivist theories of reference determination: whereas descriptivism threatens to make genuine moral disagreement impossible, reference magnetism threatens to make it impossible to represent slightly different properties. (Schroeter and Schroeter, 2013, 20)

The objection grants that reference magnetism gets certain cases right for practical terms. Thus, we can assume that **Magnetism** implies that, in ordinary Moral Twin Earth-style cases, both communities are using their moral term 'right' to refer to the same property. That is: there is one property P which is such that (i) a possible community of consequentialists who use 'right' with a moral role refer to P because P maximizes eliteness and fit with their use of 'right', and (ii) a possible community of deontologists who use 'right' with a moral role will also refer to P because P maximizes eliteness and fit with their use of 'right' . The objection then claims that this appeal to reference magnetism will get other cases wrong, namely cases where communities use 'right$_K$' and 'right$_U$' as introduced with the stipulations described above.

Magnetism is not forced to treat these cases similarly. In fact, it is highly implausible that it would. The terms 'right$_K$' and 'right$_U$' are, in their respective communities, introduced by stipulation to refer to the property that is rightness according to either the Kantian and Utilitarian theory. Thus the first predicate is stipulated by the community that uses it to refer to something like the property of avoiding violation of the autonomy of a rational agent. The second predicate is stipulated by the community that uses it to refer to the property of maximizing happiness. These stipulations constitute differences in use that are not present in the original Moral Twin Earth case. In the original case, if reference magnetism delivers the right result, the consequentialist and deontologist communities will refer to the same property in spite of using their terms differently. But the use aspect of reference-determination in Schroeter and Schroeter's case is not even close to being identical to the use in the original case.

It is true that *in one respect* the communities in Schroeter and Schroeter's case will use 'right$_K$' and 'right$_U$' in the same way as the communities in the original Moral

Twin Earth case. The users of 'right$_K$' will apply their term to all and only the same acts that a deontologist community with an undefined term 'morally right' will apply their term to. And the users of 'right$_U$' will apply their term to all and only the same acts that a consequentialist community with an undefined term 'morally right' will apply their term to. Thus the users of moral language in Schroeter and Schroeter's case will use their terms with the same individual applications as some users of moral language in the original Moral Twin Earth case.

But individual applications do not exhaust the aspects of use that are relevant to reference-determination, and so we cannot use this limited point to draw conclusions about what **Magnetism** is committed to in Schroeter and Schroeter's case. The moral terms 'right$_K$' and 'right$_U$' are used with stipulated definitions. These moral terms are stipulated to mean the same thing as a composite expression that has its reference determined by compositional rules and the reference of its constituent parts. And these constituent parts refer to whatever maximizes eliteness and fit with how *they* are used. It is extremely plausible that the defined terms 'right$_K$' and 'right$_U$' will have different referents according to **Magnetism**, because how they are used differs in an extremely crucial respect. Once we take this into account, we will not be tempted to think that **Magnetism** has implausible consequences for 'right$_K$' and 'right$_U$'.

4.2.2 Other elite candidates

Proper attention to the details of the use component in **Magnetism** avoids Schroeter and Schroeter's worry that reference magnetism will predict that too many possible practical terms will refer to the same thing. There is a mirror worry as well: that the eliteness component of **Magnetism** predicts that too few possible practical terms refer

to the same thing.

Consider a highly elite property that has nothing to do with obligation, but which is possibly in the vicinity of some community's use of normative vocabulary. One (perhaps hypothetical example) is a property that explains the evolution and adaptive value of certain traits of humans within early hominid societies. There is a property, we can assume, that individuals instantiate when they cooperate and engage in rituals that promote the survival of the group they are a part of. This is, moreover, a purely scientific fact about them, discoverable through something like the tools of evolutionary biology. We can call these traits *culturally adaptive* for short.

Another hypothetical example is the following. Suppose that, at the molecular level, some bodies contain a configuration of chemically significant molecules that enter into reactions and explain a wide range of macro-features of the bodies that contain them. This property is a particular configuration of carbon molecules fits this description. Call it *significant carbon*.

Both of these properties are—or can be imagined to be—metaphysically elite. They are stipulated not to be gerrymandered, and they play a significant explanatory role in a serious area of theoretical investigation. (Although I have not done the work to show this here, it seems straightforward to find real examples of similar properties from the higher-level and social sciences.) These are metaphysically elite properties, but are distinct from rightness or all-things-considered obligation.

If we focus simply on the individual applications of a practical term, we can imagine a possible community whose use of a practical term fits actions that are manifestations of culturally adaptive traits. They will use their term 'ought' very differently from us: in general, we do not treat the fact that an act is related to a culturally adap-

tive trait as a reason to perform the act at all. This community, on the other hand, will regularly apply 'ought' to an act of self-sacrifice that would benefit a whole early hominid community. The act is adaptive in the relevant sense. But it is not one that instantiates obligation: self-sacrifice for the good of the group is, in many cases, not required.

Likewise we can imagine a community for which significant carbon plays a similar role. Actions that involve a body containing significant carbon are, with regularity, actions this community applies their normative term 'ought' to. We, of course, think that whether an action contains significant carbon is irrelevant to whether one ought to do it. Frequently we say that one ought to do an act, when it involves no body that contains significant carbon.

It is tempting to look just at the individual applications of these terms and conclude that the use of 'ought' by these communities fits the properties of being culturally adaptive or being significant carbon fairly well. The properties are instantiated by almost all of the individual acts that these possible communities are stipulated to apply their practical terms to. It appears, from this perspective, that the first community uses their term 'ought' with a high degree of fit with the property of exemplifying a culturally adaptive trait. Likewise it appears that the second community's use of 'ought' is fit very well by significant carbon. These properties are, ex hypothesi, elite. So it would appear that **Magnetism** predicts that these communities are talking about culturally adaptive traits and significant carbon with their practical terms.

Modal claims and structural connections

The first worry for **Magnetism** as applied to practical terms relied on a conception of use that focuses only on individual applications and ignores aspects of use that involve compositional expressions. This worry threatens to ignore other important aspects of use for practical terms, which are broadly what we called the *generalizations* a community accepts involving practical terms. Here I sketch two important kinds of generaliztion.

1. Modal claims. We do not only apply our normative term 'ought' to the actions that we actually think we ought to do. In addition we engage in counterfactual thinking about normativity: roughly, what our obligations would be if things had gone differently. For instance, we can ask whether, if John had decided he would not vote in upcoming elections, we ought to have encouraged him to change his mind. This is a *counterfactual individual application* of 'ought': in answering this question we are asking whether 'ought' applies to an action in a counterfactual situation.

There are also counterfactual general principles. These cover how our obligations change (or fail to change) with systematic variations in the state of the world. Many people accept moral counterfactuals like 'if my hair had been 1 centimeter longer, it would still be wrong to cause unnecessary pain to others' or 'if people were more likely to need rescuing, we would still be required to lend aid to those in need.' These counterfactuals include general principles about unnecessary pain and the duty to rescue.

There are also counterfactual general principles that are relevant to the fit between use and elite properties like cultural adaptation and significant carbon. It is presum-

ably contingent that bodies contain carbon. Likewise evolution might have proceeded differently, or not at all. It is a contingent fact that signifiant carbon and having culturally adaptive traits are instantiated at all. Speakers in the imagined communities might still accept, or have dispositions to accept, individual counterfactual claims and general counterfactual principles about scenarios where these properties are not instantiated.

For example: speakers like us will accept the counterfactual principle 'even if there had been no significant carbon, one would still be obligated not to cause unnecessary pain to others'. Or, a community like us but whose actual individual applications of 'ought' are to actions that would be produced by culturally adaptive traits would accept the counterfactual principle 'even if nothing were culturally adaptive, one would still be obligated to give some money to charity.' (Of course this is not necessary—a point I return to below.)

2. Structural connections. Our use of normative terms is structured in certain ways. Normative obligations give rise to further obligations. Principles capturing these claims will require the candidate referents for normative terms to be distributed in the world in specific and systematic ways, if they are to be good fits with our use of normative vocabulary. One example is obligations to *intend:* if Sam ought to give money to charity, then Sam ought to intend to give money to charity. If Susan ought to go to the baseball game, and getting on the train is the only way to get to the baseball game, then she ought to get on the train.[140]

These features are structural in that they can be formulated with some degree of generality. One way to do this is to abstract away from the particular actions used in

[140]There are parallel questions about what rationality requires. Since the thin 'ought' I am using here is not (obviously) equivalent to the 'ought' of rationality, these structural requirements are not necessarily the same as those under discussion in the extensive literature on rational requirements. See Brunero (2012) and Schroeder (2009) for examples of the debate on rational requirements. And see Kolodny (2005) for one view on the relationship between the two questions.

the examples: plausibly, the following schemas capture some of the structural features of obligation:

If one ought to ϕ, and the reasons for which one ought to ϕ are moral reasons, then if one does not ϕ and does not have an excuse, one ought to be blamed for not ϕ-ing.

If one ought to ϕ, then one ought to intend to ϕ.

If one ought to ϕ and the only way to ϕ is to ψ, then one ought to ψ as well.

These structural claims, or something similar to them, capture a fact about obligation, the subject-matter we use our normative term 'ought' to talk about. As with modal claims, the communities who apply their term 'ought' to traits that are culturally adaptive, or to actions involving significant carbon may, or may not, be different in this respect. That is, they may use their term 'ought' with the structural connections listed above. Or they may use the term differently, and not respect the same structural connections we do. It is worth considering the cases separately.

3. Dependence claims. We might consider adding a further aspect of use to this list, which is essential if properties are individuated hyperintensionally. A hyperintensional individuation allows that the property of being morally right is distinct, but necessarily co-extensive with, its supervenience base. Suppose this is the property of maximizing happiness. An intensionalist individuation will hold that these properties are identical, and so if moral rightness is elite, it follows that happiness-maximization is as well. The hyperintensionalist is not committed to this.

The possibility of a hyperintensional individuation raises puzzles for reference magnetism that will not be settled by modal and structural claims alone.[141] Ex hypothesi there is a property (e.g., happiness-maximization) that is necessarily co-instantiated with rightness. So, if rightness satisfies a modal or structural claim, happiness-maximization should as well. Moreover, it is open to the hyperintensionalist to hold that *both* properties are elite. If so, **Magnetism** appears to have nothing to say about why we manage to refer to rightness with our moral term 'right'. The property it supervenes on will do equally well on considerations of fit and eliteness.

To avoid this problem, we should also focus on another aspect of use. Speakers will regularly accept claims about the dependence of the moral on its supervenience base. For example, a speaker who accepts that rightness supervenes on happiness-maximization will be disposed to say something like

If an act is morally right, its rightness depends on its being happiness-maximizing.[142]

The details of the dependence-claim will need to be fleshed out, but in a hyperintensional context the dependence-claim will serve to distinguish, for the purposes of the use-component of reference-determination in **Magnetism**, the elite moral property of rightness from the elite property it supervenes on.

Of course it is not plausible that speakers will regularly accept a dependence-claim that involves a specific necessarily co-extensive property. But that will accept that whatever elite property rightness is necessarily co-extensive with (if there is one), it is a property that rightness depends on. And this will be enough to distinguish the moral property from other properties at the level of usage. Again speakers who reject basic

[141] Thanks to a reader for raising this issue.
[142] Zangwill (2008)

platitudes like this are possible. This is an issue which should be treated separately, which I return to below.

Structure, fit, and disagreement

Begin with the assumption that the community which applies 'ought' to significant carbon embeds their term 'ought' in modal claims like we do: they say things like 'even if there were no significant carbon, it would still be that we ought to give money to charity'. Their individual applications of 'ought' to acts in their environment might routinely coincide with instantiations of the elite property of significant carbon. But significant carbon is a horrible fit with their overall use of 'ought'. There are entire worlds where no significant carbon exists, but this community is willing to say, of an act in one of these worlds, that 'ought' applies to actions in such worlds. No amount of reference magnetism can override this lack of fit.

A similar point applies if we assume that the community uses their term 'ought' with the same structural connections as us. Suppose the act of going to the grocery store is an act which instantiates significant carbon, and that this community applies their term 'ought' to going to the store (on a particular occasion). Given that they use 'ought' with the relevant structural connections, they will also apply 'ought' to intending to go to the store. If driving one's car is the only way to get to the store, then, given their use in accordance with particular the relevant connections, they will apply 'ought' to driving one's car.

There is no reason to expect that if going to the store instantiates significant carbon, then driving also instantiates the property. (We can fill in the details to make it clear in this case that this structural pattern does not hold: for instance, the act of going to the

store involves action from a carbon-based life-form; driving involves primarily a vehicle constituted by metal and so does not instantiate significant carbon, as described.) Similarly for giving to the poor and intending to give to the poor. Significant carbon does a poor job of fitting with the overall use profile of 'ought' in this community.[143]

These points apply equally well to a community whose individual applications of 'ought' track cultural adaptation. There are worlds where nothing is culturally adaptive; there are some possible worlds where everything was created 5 minutes ago with appearances that are very similar to the actual world. A community that accepts modal claims involving 'ought' resembling ours will apply the term to counterfactual constructions like this nonetheless.

Likewise there is no guarantee that a trait that promotes culturally adaptive behaviors will pattern with the structural connections of 'ought'. To take one example of a structural connection that speakers typically associate with normative terms:

If one ought to ϕ, then one ought to intend to ϕ.

Some evolutionary processes select for behaviors by hardwiring the desirable behavior into cognitive architecture. For example, reactions of laughter serve some social function but appear fairly early in cognitive development.[144] In these cases forming an intention to perform the behavior will not be adaptive, since the behavior will occur without the intention. One doesn't need to intend to laugh in order to do so. Culturally adaptive acts are not always accompanied by a culturally adaptive intention.

[143]We can combine the point involving modal claims with the structural point. Even if significant carbon happens to fit the structural profile of 'ought', this will be the result of a contingent accident. There are other worlds where significant carbon does not fit the profile. A community that endorses modal claims like ours will endorse the claim that 'ought' obeys the structural connections even in worlds where significant carbon does not.

[144]See, for example, the overview of theories of humor in Olin (2016).

These arguments rest on the details of a few examples. But they show something important that plausibly generalizes across other cases. The modal and structural claims we make involving practical terms are important aspects of how we use these terms. Use is not exhausted by the individual acts that a community applies their practical terms to. Once we take seriously the modal and structural claims communities make using their practical terms, it becomes very difficult to find an elite subject-matter from a scientific discipline that fits our use of practical terms well.

There is a second point to make about these examples. We can imagine possible communities that use their practical terms without accepting these modal and structural claims. We should say something very different about the consequences of **Magnetism** for the reference of practical terms as used by these communities.

The discussion of the **Universal Disagreement** thesis in Chapters 1 and 2, we focused on *every* possible community that uses moral or normative terms. To use an expression as a normative term requires using the term with a normative role. Likewise, to use an expression as a moral term requires using the term with a moral role. The argument against **Universal Disagreement**, in outline, is that there are possible communities that use an expression with a moral or normative role, but who also use their term with (or without) additional roles that make it very natural to think that they are not talking about the same thing we are talking about with our practical terms.

A use of 'ought' as a normative term does not necessitate accepting the modal and structural claims outlined above. Some possible community coherently uses their term 'ought' with the Gibbard role, while simultaneously rejecting, or at least not accepting, the modal claim that worlds which are very different from ours in empirical respects contain obligatory acts. Other possible communities deviate in other ways from the

modal and structural claims we accept. Such possible communities do not provide the same grounds for holding that, according to **Magnetism**, 'ought' in their mouths does not refer to cultural adaptiveness or significant carbon. Since these communities do not accept typical modal and structural claims they will, in principle, allow these properties to fit much better with other elite properties.

Since **Universal Stability** is false, there is no guarantee that, just because a community uses a term with a practical role, they are thereby referring to what we are referring to with our practical terms. We have seen in Chapter 2 that a shared thin practical role, whether it is the Gibbard role or the moral rightness role, does not guarantee co-reference. Some possible communities use their practical terms with additional roles, which make it clear that they are not referring to what we are referring to. On some possible ways of filling in the additional roles a community uses their 'ought' with, communities that share a practical role with us are not referring to the same property. We need the shared roles between communities that have the same practical terms to be *thicker*, before it is intuitively clear that they can disagree with each other.[145]

We should be willing to treat communities that do not use their practical terms with the same thick practical role, by dropping certain central modal or structural claims from their use, as an instance of the same phenomenon. The communities in question use 'ought' by, for example, rejecting some instances of claims structural or modal claims such as the following:

If one ought to ϕ, then one ought to intend to ϕ.

[145]Thick roles should not be confused with another use of similar terminology, on which certain moral *terms* are "thick". Here I just mean that a term used with the Gibbard role, for instance, needn't be used in the modal or structural claims sketched above. See Roberts (2011) for more on this alternative notion.

If evolution had gone differently (or had not occurred at all), it would still be the case that we ought to ϕ.

It is not at all intuitively obvious that such a community would be using their normative term 'ought' to be referring to to obligation. Moreover, since we already know that **Universal Stability** is false, we should not take the mere stipulation that these communities are using their terms with the same practical role to show that **Magnetism** must entail on pain of explanatory inadequacy that such communities are disagreeing. Divergent individual applications to non-normative elite properties, that are accompanied by a rejection of a thick practical role, are further counterexamples. It is not implausible to interpret such a community as using 'ought' as a normative term but not referring to obligation.

On the other hand, if these communities are also using their terms with the same thick role, then **Magnetism** should, and has the resources to, predict they are talking about the same thing, in spite of divergent individual applications. In these cases, properties like being culturally adaptive, or being significant carbon, are extremely poor fits with the modal and structural claims that make the practical role thick.

Thus, possible communities that do not use their practical terms with the same thick role as us are not counterexamples to **Magnetism**. But someone with an antecedent commitment to **Universal Stability** might think that they are. Williams (2018) presents an object to reference magnetism for practical terms, which appears to be motivated by this assumption.

Williams's objection

Williams (2018) presents an objection along the lines sketched above, although it

raises additional complications. But the response to the Williams objection will in outline be the same as that outlined above: he claims (i) that it is easy to imagine communities that use their term 'ought' to fit with highly elite properties from the sciences, and (ii) that if a community does this, they should still come out as talking about the same thing we are talking about. Reference magnetism, Williams claims, fails to explain this.

Here is a passage where Williams is discussing the use of reference magnetism in Dunaway and McPherson (2016):

> [S]uppose that E is a community who have a concept that plays the internal conceptual role of wrongness, and who take this to pick out a property P where (perhaps unbeknownst to them) P plays a basic role in some *other* serious science. Perhaps, for example, *maximizing self-interest* is a property that has a basic explanatory role elsewhere in normative theory (or economic theory, or psychological theory...); and E are a community of egoists, thinking that the right action (for x) is that which maximizes x's self-interest. In virtue of its explanatory role outside morality, the property onto which E have latched will be maximally eligible—just as eligible as authentic moral permissibility. In virtue of their egoist moral theory, interpreting "morally wrong" as *failing to maximize self-interest* will score better than rivals in terms of maximizing truths attributed, and is no worse on the dimension of eligibility. So here we have a case where a community has a concept that plays the internal role of moral wrongness, but fails to thereby denote moral wrongness itself. That is exactly what we needed to avoid. (Williams, 2018, 57, his italics)

First, there is an important distinction to make about this example, which does not arise in the simple case sketched earlier. Williams specifies that self-interest maximization is "a property that has a basic explanatory role elsewhere in normative theory (or economic theory, or psychological theory...)" and hence is elite. But the type of elite property that self-interest maximization (allegedly) is matters here. If it has a role in *normative* theory, then in light of the failure of **Universal Stability**, we shouldn't necessarily assume that there is something wrong with interpreting a community that uses

their normative term 'wrong' in this way as referring to something other than moral wrongness. Perhaps they use the term with the "self-interest role" and constitute another Twin Earth community that uses their normative term to talk about something different than what Earthlings talk about. They are not speaking about wrongness, but about some other normatively relevant property. They would be another counterexample to **Universal Disagreement** that fits the template in Chapter 2.

We do need to concede that even though **Universal Stability** is false, moral and normative terms are highly stable, and so a good meta-semantic theory will need to explain why many linguistic communities across modal space use their practical terms to refer to the same property. It is not in general true, then, that we can respond to a Williams-style counterexample by conceding that the communities in question are referring to different properties. We will need to look carefully at the detailed descriptions of the alleged communities on a case-by-case basis, to determine whether it is plausible to hold that they are talking about distinct normatively relevant properties. But: Williams is aiming for a schematic example which shows, in his words, that reference magnetism suffers from a "structural defect".[146] The argument as presented fails to deliver this result, when the candidate referent (self-interest maximization) is stipulated to be an elite normative property. This version of the objection rests on the false assumption of **Universal Stability**. In fact in light of the failure of **Universal Stability**, a structural counterexample of the kind Williams aims for is impossible: we need to rely on the concrete particulars of a case to determine whether it is a case where speakers should be interpreted as referring to the usual properties. This cannot be read off of the roles that are stipulated to be a part of their usage alone.

[146] Williams (2018, 57)

On the other hand Williams also suggests a different, and potentially more worrying, objection. Suppose self-interest maximization is not an elite property that is relevant to normative theorizing, but rather is elite because it plays a fundamental role in psychological explanations of human behavior. If reference is determined by maximization of use and eliteness, as **Magnetism** holds, then a community that applies their normative term 'ought' to enough self-interest-maximizing actions will, according to Williams's objection, refer to self-interest-maximization. But this will generate new, implausible claims about instances of possible communities who fail to have substantive disputes with each other. These are cases where there is no other plausible normative subject-matter for the alternative community, and so are not the kind of cases where **Universal Stability** fails.

For example: take the community in *Selfish Twin Earth*, where speakers use a normative term 'ought' with the Gibbard role: their applications of 'ought' to an action guide action. But the Selfish Twins routinely apply their term 'ought' to self-interest-maximizing actions, saying things like the following

John ought to defect from a prisoner's dilemma;

Jamie ought to spend almost all of her money on things she wants and ignore the needs of others.

Ex hypothesi, self-interest maximization is instantiated by the actions they apply 'ought' to, including defecting from prisoner's dilemmas and spending one's money selfishly; moreover, this property of self-interest maximization is not normatively interesting. So this should not be a case that constitutes a failure of **Universal Stability**; rather, the Selfish Twins should be having a substantive disagreement with users of normative language on Earth, who say

Everyone ought not to defect from a prisoner's dilemma;

One shouldn't spend almost all of one's money on things one wants and ignore the needs of others.

According to Williams, reference magnetism entails that that the Selfish Twins are *not* having a substantive disagreement. Self-interest maximization, as a property that is significant for psychological explanations, is just as elite as the (distinct) property of obligation. Since the Selfish Twins regularly apply their term 'ought' to self-interest maximizing acts, many of which are not obligatory, their usage of 'ought' appears to fit self-interest maximization better than obligation. Since these properties are equally elite, reference magnetism implausibly implies that the Selfish Twins are not referring to obligation.

This argument is convincing only if *all* aspects of usage of 'ought' by the Selfish Twins are fit better by self-interest maximization. As we have seen, use is constituted by more than their individual applications of the term. Modal and structural claims are important aspects of usage as well, and the assumption that self-interest maximization is an elite property from psychological theory suggests that it be an extremely poor fit with use of a normative 'ought'.

Begin with modal claims. Assuming that self-interest maximization is an elite psychological property, the psychological role that it plays is contingent. Human psychology (or psychology for agents that are very much like humans) could easily have been different. If self-interest maximization plays a central explanatory role in actual human psychology, perhaps by explaining why humans regularly take actions that maximize self-interest, then there are possible worlds where humans are not self-interested, and human action is explained by other properties. There is no elite property of self-interest

maximization in the psychological sense in these worlds. Call such a world an *altruistic world*.

If the Selfish Twins use their normative term 'ought' with the normal modal claims in addition to the Gibbard role, then they will apply their term 'ought' to some actions in the altruistic world. They will make assertions along the following lines:

> Even if he were in the altruistic world, John ought to defect in a prisoner's dilemma;
>
> Even if she were in the altruistic world, Jamie ought to spend almost all of her money on things she wants and ignore the needs of others.

But there is no elite psychological property of self-interest maximization in the altruistic world, so assigning this property as the referent of 'ought' fits very badly with the use of 'ought' in these modal claims. And this is just one example: there are many other worlds that play the same role as the altruistic world. Even the Selfish Twins are not plausibly interpreted as speaking about self-interest maximization, under these assumptions.

A similar point applies to structural claims. As a psychological property, self-interest maximization fares poorly on this count as well. Take the structural claim

> If one ought to ϕ, then one ought to intend to ϕ.

It is very implausible that, as a claim about an explanatory psychological property, any time an action has the self-interest maximization property, the intention to do so has the same property. Explanations of action are the target of one area of psychological theorizing—"behavioral psychology"—whereas folk psychological notions like intention do very little explanatory work in explanations of the inner workings of the brain. As with modal claims, the structural features of a thick normative term appear to

make psychological properties (or any properties from the special sciences) candidate referents on the basis of their fit with use, even by the Selfish Twins.

Here it is also worth rehearsing a lesson from the discussion of simplified cases involving properties like significant carbon. Of course we can imagine a variant of the Selfish Twins—call them the *Thin Selfish Twins* where individual applications of 'ought' are exclusively applied to self-interest maximizing actions, but where the normative role is thin instead. The Thin Selfish Twins use their term 'ought' with the Gibbard role but none of the modal and structural claims that English speakers associate with normative terms.

In such cases we should not automatically concede that the Thin Selfish Twins, so described, are talking about all-things-considered obligation. *If* **Universal Stability** were true, it would follow simply from the fact that the Thin Selfish Twins use 'ought' with the Gibbard role that they are referring to this property. But **Universal Stability** is false. Moreover, the Thin Selfish Twin resemble the kinds of possible linguistic communities that constitute counterexamples to **Universal Stability**, since they meet the minimum conditions for possessing a practical term, but differ substantially in the additional roles they use the term with. Such communities are unlikely to supply decisive counterexamples to **Magnetism**.

4.3 The positive picture: reference magnetism, robust stability, and alternative practical subject-matters

Reference magnetism, as a theory about what determines the reference of terms in a language, can be motivated and defended both on general grounds, and as a theory of practical terms. In this section I apply the theory to the apparent facts about practical

disagreement.

There are two points on which presentation will be deliberately incomplete. First, in keeping with the primitivist theory of eliteness sketched here, I do not argue for, or motivate, the claims I make about which properties are elite. Instead, I simply assume that some specific properties are elite, and show that the theory explains the contours of the data outlined in Chapters 1 and 2. I the next chapter I discharge the assumption, by arguing that even though these claims about eliteness cannot be explained in further terms, they can be shown not to be arbitrary, or ad hoc stipulations.

Second, I will work with a somewhat schematic picture of which possible communities are not talking about the same subject-matter as users of practical language in English. Chapter 2 gave three examples of such communities: those that use their practical terms without the ability role, those that use their practical terms with the psychological feasibility role, and those that use their terms with the bounded obligation role. These are just examples; it is entirely possible that there are other structurally similar cases that are not mentioned here.

Here is a brief recap of what needs to be explained. Familiar uses of Moral Twin Earth cases suggest a general claim, which we called **Universal Disagreement**. This is the claim that every possible community that uses their terms with the same practical role is talking about the same property. While the generalization to **Universal Disagreement** is not warranted, the Moral Twin Earth thought experiment does point, for the realist, to the *robust* stability of practical terms. Possible communities who apply their practical terms in accordance with different substantive theories should be interpreted as speaking about the same thing, and hence capable of having a substantive disagreement with each other. But there are limits to this thesis—there are some ways

in which possible linguistic communities can use a practical term of the same type, but are not plausibly interpreted as speaking about the same thing.

I will begin by using **Magnetism** to explain the robustness of the stability of practical terms. Then I will turn to showing how **Magnetism** can also explain the limits to the stability of practical terms.

4.3.1 Robust stability

The first step is to explain why, in canonical cases like Horgan and Timmons's Moral Twin Earth case and nearby variants, the communities in question refer to the same property with their practical terms, and hence are capable of having a substantive disagreement. These are the cases where the substantive theory that guides application of practical terms in each community is different. We can begin with the assumption, for simplicity, that one of the communities in the case regularly applies their moral terms to elite moral properties. Suppose, for instance, maximizing happiness is the elite property of moral rightness, and so the consequentialist community applies their term 'right' to acts that instantiate the elite property of maximizing happiness. **Magnetism** straightforwardly implies that they are referring to moral rightness.

This is not simply because their individual applications of 'right' are to morally right actions. There are other aspects of their usage that are fit quite well by a highly elite property. Ex hypothesi, their term 'right' is a moral term, so they use it with the moral rightness role: whenever they apply 'right' to an action, they feel guilt for not performing it themselves, or blame others who do not perform it. And there is an elite property that fits this role: moral rightness, which is (we are assuming for illustrative purposes) the property of happiness-maximization. That is, since it is true

that blame is warranted when one does not perform a happiness-maximizing act (and similarly for blaming others in the same situations), their use of 'right' fits happiness-maximization. The moral role of 'right' as used by the community in question contributes to the fit with happiness-maximization which, ex hypothesi, is the property of moral rightness.[147] Moreover happiness-maximization provides a decent fit with other modal and structural claims that normal users of moral language will accept.

Since we are assuming, for illustrative purposes, that happiness-maximization is the elite property of moral rightness, it does extremely well on the two components of reference-determination for the moral term 'right' according to **Magnetism**. The consequentialist community uses 'right', both in virtue of their individual applications of 'right', but also in virtue of the moral and other roles they use the term with, to fit happiness-maximization very well. Since happiness-maximization (we are supposing) is elite, the consequentialist community refers to moral rightness, under these assumptions.

To explain why practical terms are semantically stable, we need to explain why other possible communities would be referring to the same property with their term 'morally right'. For example, we need to explain why it is the referent of 'right' in the mouths of the community of deontologists in the original Moral Twin Earth scenario. **Magnetism** can do this, because (i) the same property, namely happiness-maximization, is elite in other possible worlds, and (ii) the use of 'right' by the deontologist community is fit by the same property fairly well. While not all of the individual applications of 'right' by the deontologists will fit happiness-maximization, many of them will. Moreover, they use their 'right' with the same moral role and accept the

[147] Cf. Wedgwood (2001)

same structural claims as the consequentialist community. **Magnetism** only requires that the referent of a term *maximize* the two components of reference-determination. The degree of fit of 'right', in the mouths of the deontologists, is not as high as the fit of 'right' in the mouths of the consequentialists. But so long as properties that fit *better* than rightness are not elite, the referent of 'right' in the mouths of the two communities will be the same. So, **Magnetism** already can claim to explain a measure of stability in the moral term 'right'.

None of this hinges on the specific assumption that happiness-maximization is elite. Perhaps the property of doing the most good without violating the autonomy of a rational agent is metaphysically elite. Or perhaps neither community has it right: there is some third property, characterized by the correct first-order moral theory, which is neither Utilitarianism nor deontology and identifies rightness. The lesson will be the same. Both communities use their term 'right', at worst, with individual applications to actions that do not instantiate the elite property of moral rightness. But fit will still be pretty good: they use 'right' with a moral role that is perfectly fit by an elite property, and not other candidates. And, if they are communities that are clearly talking about moral rightness, their use will fit the property in virtue of their use of 'right' in various modal and structural claims. For each such possible community with a "thick" term 'right' that is used with a moral role, **Magnetism** will say: rightness is the referent, because it is both elite and fits usage of the community in question pretty well. The fit with individual applications is moderately good, and is excellent with modal and structural claims.

To give this explanation, I have been making assumption that there is an elite property of moral rightness that is in the vicinity of the usage of 'right' by communities

that use it as a moral term. This assumption will be defended in Chapter 5. But it also requires a second, negative assumption: that there is not *more than one* property in the vicinity.[148] This assumption will also have to be justified. For now, we can simply note that the outline of the explanation depends on the absence of a multiplicity of elite properties: if both happiness maximization and the absence of autonomy violation were elite moral properties, then the explanation of stability across the Moral Twin Earth communities would fail.

Nearby variants on the original Moral Twin Earth case can be explained in the same way. There are scenarios where the communities accept, and apply their moral terms in accordance with, first-order moral theories that are not simple versions of Utilitarianism and deontology. For similar reasons they will be talking about the same highly elite property of moral rightness. Communities that are not unanimous in accepting a single moral theory (such as actual users of moral language) will be even better candidates for users of a language that refers to a highly elite moral property. Since there is no unanimity in such a community, there is no single property that best fits the individual applications of their terms. But they will agree on 'right' as a moral term, and be near-unanimous on the association of 'right' with the moral rightness role, and other modal and structural claims involving 'right'. Moral rightness will fare no worse on the use component of reference-determination according to **Magnetism**, and will fare better on the eliteness component. **Magnetism** can explain a significant amount of stability for moral terms.

A similar story applies to bare normative terms. Take two possible communities that use their normative terms with the Gibbard role, but differ in which substantive

[148] Dunaway and McPherson (2016) call this assumption "uniqueness".

normative theory they accept. Assuming that there is an elite normative property in the vicinity, **Magnetism** can hold that both communities will be talking about it. Both communities will use 'ought' with a role, along with other modal and structural claims, that are fit very well by this property. Moreover, there are no other elite properties in the vicinity that likewise fit the role decently well. This explains why, as **Robust Stability** claims, shared practical roles guarantees shared reference across a wide variety of linguistic communities across modal space.

4.3.2 Limits to stability

Not all possible linguistic communities use their practical terms to refer to the same properties that we refer to. The clearest examples of these communities are those that use 'right' or 'ought' the requisite moral or normative roles, but differ from typical English speakers by using these terms with additional roles that are not a part of normal English usage. For example, there is a possible community that uses normative terms without the *ability role*, as they systematically ignore whether the actions they apply 'ought' to are actions that the relevant agent has the ability to perform. Their standards otherwise appear to be like ours. These are the Ability Twinsfrom Chapter 2.

The Ability Twins appear to be talking about something that is not obligation. It is obligatory for people with disposable income to give some of their money to charity. It is not obligatory for such people to end an entire famine. But the Ability Twins routinely apply their term 'ought' to an action that would create the best consequences simpliciter, including ending a famine. They do not take it as a relevant consideration that the subject of the 'ought' is able, in any normal sense, to perform the action. This

is not within the abilities of Jamie. The Ability Twins are not, intuitively, talking about what Jamie is obligated to do, in the normal sense. They are talking about what would be the best state for the world to be in, regardless of Jamie (or anyone else's) contingent limitations. This is the *best state property*.

Since the Ability Twins are one clear counterexample to **Universal Stability**—and, by extension, **Universal Disagreement**, since we would not treat ourselves as having disagreements with them—I will focus on showing how **Magnetism** explains this. I will focus here on the Ability Twins, and will not provide an entire catalogue of the counterexamples to **Universal Disagreement**. Developing an explanation of the referent of practical terms as used by the Ability Twins should provide a model for extending the explanation to other cases.

It is not enough simply to use **Magnetism** to explain why the Ability Twins are referring to some property distinct from obligation. Of course this is one necessary part of a good explanation—we need to explain why **Universal Disagreement** fails in cases involving the Ability Twins. But, since we have appealed to **Magnetism** to explain why practical terms are highly stable, we need to do more: we need the **Magnetism**-centric explanations to be consistent with one another. If we can do this, we will have the beginnings of an explanation of why **Universal Disagreement** is false, but **Robust Disagreement** is true.

Begin with the assumption that the best state property is elite, and hence is a reference magnet. This is a distinct property from the property of obligation simpliciter, which is also elite. Relieving a famine has the best state property, but Jamie is not obligated to do it. Regardless of whether the Ability Twins are always successful in applying their term 'ought' to the best state property, we should imagine the case

to be one where the best state property fits their individual applications of 'ought' reasonably well.

There are other dimensions along which the best state property fits the Ability Twins' use of 'ought'. The best state property will fit fairly well with the usual modal principles that are associated with normative language. For example, when speaking about counterfactual situations, no matter how distant, the Ability Twins will apply 'ought' to the actions that instantiate the best state property in those situations. The best state property is not like significant carbon, which is not instantiated in some possible worlds that resemble ours in normative respects. In addition, the Ability Twins use their 'ought' without the ability role—they reject claims such as 'if Jamie ought to relieve the famine, then she is able to relieve the famine'. The best state property fits this role for 'ought' in the Ability Twins' mouths very well.

The best state property is not a perfect fit. The Ability Twins use their term 'ought' with the Gibbard role. It is not inconsistent for a community to use a term with both the Gibbard role and with the absence of the ability role. But the best state property is not a good fit with the Gibbard role; instead, the property of obligation is. It is not incoherent to think that some act has the best state property and not perform that act, but it is incoherent to think that an object is obligatory and not perform it. Since the Ability Twins use 'ought' with the Gibbard role, they treat whatever property 'ought' refers to as one which fits the Gibbard role. But, given the total overall profile of their use of 'ought', there is no property that perfectly fits the Ability Twins' use of 'ought'.

Magnetism can accommodate this, since reference is determined by maximizing fit with *overall* usage, plus eliteness of referent. A proponent of **Universal Disagreement** holds, in effect, that if 'ought' is used as a normative term, its referent is the property

that best fits the normative role of 'ought'. But **Magnetism** does not say this. If there are other roles that 'ought' is used with, the property that best fits the use of 'ought' by a community is one that best fits all of these roles. Given that other aspects of the Ability Twins' use of 'ought', including their individual applications and the absence of the ability role, are perfectly fit by the best state property, it is very plausible that the best state property is a better fits their use of 'ought' than obligation—even though 'ought' is used with a normative role by the Ability Twins.

Finally, we can complete the explanation: both the best state property and obligation are elite. Since the best state property is, for the reasons sketched above, a better fit with the Ability Twins' use of 'ought' than obligation, **Magnetism** implies that they refer to the best state property.[149]

We also need to show that this explanation is compatible with significant semantic stability for practical language. For communities that use the normative 'ought' *with* the ability role, including normal English-speakers but not the Ability Twins, the best state property is a horrible fit with their use. Even though it is elite, it is a poor candidate referent for the normative 'ought' as used by these speakers, since it fares so poorly on the use component of reference-determination. Thus, among the properties that are candidate referents—that is, among those that are at least pretty good fits with the use of 'ought' by normal speakers—only one of these properties is elite. That property is obligation. A normative 'ought' as used by these speakers will be highly stable, which is what the explanation of **Robust Stability** requires. Appealing

[149]What about a community like the Ability Twins that systematically applies 'ought' to acts that would not bring about the state that is in fact the best one for the world to be in (perhaps because they have false moral views)? In that case neither obligation nor the best state property are perfect fits with individual applications. The best state property fits the absence of the ability role; obligation fits the normative role. So which property are they referring to? It is not obvious that they will succeed in referring to the best state property according to **Magnetism** (perhaps it implies that reference is indeterminate in a case like this). This is not necessarily a bad result: intuitions about cases like this are unlikely to be very strong.

to **Magnetism** to explain why the Ability Twins do not refer to obligation does not jeopardize its explanation of the semantic stability of 'ought' in other cases.

This completes the sketch of why **Magnetism** explains the failure of **Universal Stability** in the Ability Twins case. Of course, if the arguments from Chapter 2 are correct, there are other cases where users of normative language are not referring to obligation. Instead of walking through the details of an application of **Magnetism** to these cases, I will simply note that the explanation proceeds in roughly the same way, given the assumptions that the property of *bounded obligation*, and the property of *psychologically feasible obligation*, are elite. When a possible linguistic community uses their normative 'ought' with additional roles that fit these properties fairly well, we get further cases where **Magnetism** implies that users of a normative 'ought' are not referring to obligation.

The upshot for practical language is that there are multiple possible subject-matters for moral and normative terms. Speakers of languages that pick out different referents with their practical terms are talking past one another, and are capable of having only merely verbal disputes. I will return in concluding remarks to a discussion of the implication of these assumptions for realism. But it is worth noting how at a broad level none of this is inconsistent with realism about morality and normativity. Instead it follows from a straightforwardly realist approach to the metaphysics of morality and normativity, and the meta-semantics of practical language. Some moral properties are elite; moral language describes the properties that best maximize fit with the language and eliteness. The same goes for normative language: it describes an elite subject-matter of normative properties. The reason why **Universal Disagreement** fails is that there are multiple moral and normative joints, each of which is highly elite.

4.3.3 Addendum: holism and the contribution of eligibility to practical role

The foregoing presents the contribution of eliteness to reference-determination by engaging in the pretense that the referent of a practical term is solely determined by the degree of fit of a referent with the use of that term, plus the eligibility of candidate referents. This is, officially, not true, since reference determination is, as I have noted, a holistic matter. The referent of a determined by which interpretation maximizes fit and eliteness across an entire language.

In most cases ignoring the holistic character of reference-determination is harmless, as focusing on the eligibility and fit of candidate referents will be a good heuristic for determining what that term refers to. But there is one respect in which the foregoing discussion relies on the holistic approach, and this is worth highlighting. This can be seen by highlighting two claims that the explanation of stability in terms of reference magnetism relies on. These are:

Fit with role Rightness fits the use of 'right' with the moral rightness role.

Eligibility Rightness is elite.

According to **Magnetism**, both **Fit with role** and **Eligibility** are part of the reference-determining facts for the normative 'right'. Any possible linguistic community that uses 'right' with a moral role will, to that extent, use 'right' with a degree of fit with rightness. It is the property of rightness that explains why blaming or feeling guilt for certain actions is warranted. When a community uses 'right' with a moral role, this aspect of their use is thereby a better fit with rightness than with other properties that do not warrant guilt and blame. Their overall pattern of usage might not be a perfect fit with rightness, as communities who adhere to false moral views will not

use their 'right' to fit perfectly with rightness, by applying the term to actions that are not right. Some communities like this nonetheless succeed in referring to obligation, and according to **Magnetism Fit with role** partly explains this fact.

Magnetism also holds that **Eligibility** partly explains why such communities refer to obligation. The reason why 'ought' refers to obligation, rather than some other property that is less elite but fits better with community-wide use, is that obligation is elite while other candidate referents are not.

What this leaves unclear is why, given that rightness does better than other properties in the way **Fit with role** describes, a theory of reference-determination also needs **Eligibility** to explain why 'right' refers to moral rightness. If we are already spotting a theory of reference-determination the claim that rightness is the property that best fits with the moral role of 'right', then in principle a meta-semantic theory could hold that this suffices for reference to rightness.[150] The metaphysics of role I have given here implies that the moral role of 'right' is just one part of the use of the term, and so is not capable of settling reference on its own. But there is also a deeper reason why treating **Fit with role** as sufficient for reference to rightness does not show that **Eligibility** plays no role in reference-determination.

The moral rightness role is one of a wide range of available roles. Not only are there roles that have very different flavors—a normative role, or non-practical role, that we might use a term with—there are also roles that resemble the moral rightness role, and differ from it only in virtue of being connected to gerrymandered neighbors of guilt, blame, and the like. For example there is the activity of *blaming**, which one blame*s someone just in case one does something that is a lot like blaming, but does

[150]Cf. Wedgwood (2001)

not occur at 9am on Sundays that fall on odd-numbered days before 1000AD. There is an analogously gerrymandered activity of feeling *guilt**. And moreover there is a relation of appropriateness* which holds between some acts and guilt* and blame*. A moral wrongness* role can then be defined as follows:

> A term *t* is used by a community with the moral wrongness* role just in case speakers treat it as appropriate* to blame* other agents who perform actions that the community applies *t* to, and appropriate* to feel guilt* when they perform actions the community applies *t* to.

The moral rightness* role is then the complementary role to the moral wrongness* role.[151]

Even if it were true that *Fit with role* suffices to determine the reference for 'right', we need to explain why it is that it is the moral rightness role, rather than the moral rightness* role, that determines reference. This is the contribution of holism: an interpretation of a community should maximize fit and eliteness for all terms of a community's language simultaneously, and not just individually. In particular, it should treat a community as speaking about guilt and blame, rather than guilt* and blame* or some other gerrymanders, when they use their term 'right'. Interpreting speakers as using their terms with a moral role rather than a moral* role will significantly increase the overall eliteness of the referents in their language, since it treats them as referring not only to rightness (which we are assuming is elite), but also to guilt rather than guilt*, and blame rather than blame*. The holistic aspect of reference-determination thus suggests that **Eligibility**, and related facts about the eliteness of the properties referred to

[151] We can even define a *consequentialist blame*, which is like blame, but applies only to acts that are wrong by consequentialist standards. *Consequentialist guilt* works similarly, and given a notion of *consequentialist appropriateness* there is a consequentialist moral wrongness role that is fit by consequentialist users of moral language very well.

by terms that constitute a practical role, are an essential part of a **Magnetism**-based meta-semantics for practical terms.

Of course, as I have argued in the foregoing, a meta-semantics that relies only on **Fit with use** will be inadequate for other reasons, since it will overgeneralize. What the holstic aspect of reference-determination adds is that, even when speakers are referring to the property that best fits the practical role they associate with their moral or normative terms, **Eligibility** still makes an important explanatory contribution to why this matters for reference.

4.4 Contextualism with elite rankings

So far I have articulated reference magnetism under the assumption that the way to explicate the meanings of practical terms is by specifying the property that they refer to. This is inessential to the picture, though making the same points outside this simple semantic framework is not straightforward. Here I note some of the most salient features of an implementation of the same view which assumes a contextualist account of practical terms along the lines of Kratzer (1977), which we outlined in earlier chapters.

Take 'ought', used (on a particular occasion) as a moral term in the following sentence.

You ought to apologize.

If this sentence is (as used on the relevant occasion) true, then it is because the act of apologizing' ranks higher according to the relevant moral standards (the "ordering source") than any other alternative act (in the "modal base"). The ranking is simply a relation among acts: in a simple version of this case the relevant acts are the act of

apologizing and the act of not saying anything.[152] If some properties are elite, then it is natural to extend this idea by claiming that the ranking, which is a relation between apologizing and not saying anything, is also elite. Properties are simply relations with only one argument place. If they can be elite, then it is a natural step to hold that relations with two or more argument places can be elite as well.

In some contexts, speakers use 'ought' to rank actions according to a highly elite ranking. The contextual flexibility of 'ought' does not always pick up on such a ranking: for instance when intentionally evaluating actions against what would make them happiest, two speakers, *A* and *B* might each truly say the following of themselves:

A : I should sit on the couch all day;

B : I should go on a ten-mile run.

Assuming that *A* is made most happy by sitting on the couch, and *B* enjoys long runs, each says something true.

But sometimes speakers' intentions to speak about what morality requires expresses a claim about a moral ranking. Explaining these uses as capable of featuring in substantive disagreements with other requires that the thesis we called **Ranking Stability** in Chapter 2 is true:

Ranking Stability Any possible who use their term 'ought' with the same moral or normative flavor are thereby making claims whose truth-conditions are determined by a single ordering source.

Ranking Stability captures why an analogous dispute about morality does not result in each speaker saying something true, and hence failing to substantively disagree.

[152]Schroeder (2011). On alternative views, 'ought' is an operator that applies to complete sentences (Wedgwood, 2007). In this case it is a property that applies to propositions; the distinction will not be important for what follows.

Analogues of *A* and *B*—call them *C* and *D*—disagree about what morality requires when they intend to use moral rankings when asserting the following:

C : Jamie ought to give at least $500 to charity;

D : It is not the case that Jamie ought to give at least $500 to charity.

Ranking Stability says that *C* and *D* are disagreeing because there is a particular moral ranking of actions, such that *C* is claiming that Jamie's giving at least $500 to charity ranks highest according to that ranking, and *D* is claiming that giving giving $500 does not rank highest according to *that very same ranking*. They are having a substantive dispute.[153]

Importantly, the presence of a substantive disagreement persists even if *C* and *D* explicitly accept different claims about how the moral ordering source which determines the truth-conditions for their assertion, ranks the action of giving $500 to charity. For instance, *C* might think that morally requires maximizing happiness, and so Jamie's giving at least $500 ranks highest against the relevant alternatives because it produces the most happiness. Call this the *happiness* ranking. *D*, on the other hand, thinks that morality only requires not violating the rights of others. Since not giving $500 violates no one's rights, it is among the highest-ranking actions according to moral standards. Call this the *rights-violation* ranking.

Even though *C* and *D* intend to make claims about the happiness and rights-violation rankings, respectively, at least one of them fails. They are both using 'ought'

[153]This claim needs several qualifications, which I will assume are in place throughout. First, the disputants do not use their moral 'ought' in the relevant context with (or without) any of the additional roles from Chapter 2. If they do, this will not call for an account of their dispute that characterizes it as a substantive disagreement. Second, they are not modally separated in worlds that differ factually in morally relevant ways. For instance, if *C* is in a world like ours, where Jamie makes a healthy income, while *D* is in a world that differs in that Jamie is very poor, then there is no pressure to interpret them as substantively disagreeing. This latter qualification is the same as the qualification on the notion of a substantive disagreement earlier. Even if two speakers accept logically inconsistent claims, this is not sufficient for substantive disagreement.

with the same moral flavor, and hence, according to **Ranking Stability**, are making claims that have their truth-conditions determined by the same ranking. At least one speaks falsely. Reference magnetism can explain why (although calling the meta-semantic thesis *reference* magnetism is a misnomer in this case). Given each speaker's intention to evaluate Jamie's act morally, eliteness considerations select a particular moral ordering source for the truth-conditions of their assertions. Which ranking is relevant is no doubt in part determined by the fact that they intend to rank actions morally. But, just as **Magnetism** holds that use is not the only component of reference-determination, an extension of the picture to the contextualist semantics for 'ought' should hold that the eliteness of a particular ranking is also relevant for which ranking determines the truth-conditions of a particular assertion involving a moral 'ought'. This may not be the ordering source that the speakers intend, since one intends to use the happiness ranking, and the other intends the rights-violation ranking. But none of this is a substantial departure from what we are already committed to un a simple semantic framework, where practical terms refer to properties. Some communities intend to refer to happiness-maximization with their moral 'ought', and fail to do so because morality does not always require happiness-maximization. The contextualist picture, for a realist, will need to acknowledge a similar phenomenon.

Given the failure of **Universal Stability** in a simple semantic framework, we will need to add some complications to the contextualist view in order to avoid overgenerating predictions of moral and normative disagreement. Someone who is using their term 'ought' with the bounded optimality role will make claims about which acts rank best according to one elite ranking, namely a ranking that places acts which are boundedly optimal highest. They will not disagree with someone who uses 'ought' with a

purely moral flavor; it is not inconsistent to claim that ϕ-ing ranks highest according to the bounded optimality ranking and does not rank highest according to moral standards. These are different rankings. There is no substantive disagreement in cases where speakers make claims about where the act of giving $500 ranks on different ordering sources.

Other cases where **Universal Stability** fails do not fit this pattern in one important respect. Take someone who uses a moral 'ought' without the ability role. They are not disagreeing with someone who uses their moral 'ought' with the ability role. For instance a speaker using 'ought' without the ability role can assert 'John ought to end the famine' truly; someone else using the ordinary moral 'ought' will truly say 'it is not the case that John ought to end the famine'. There is no disagreement here, but this is not because the speakers are making claims about different ordering sources. Instead the *modal base* is different. The speaker who uses 'ought' without the ability role is claiming that, out of all the actions someone without contingent limitations could perform, ending the famine is among the best. Someone who uses 'ought' with the ability role is claiming that ending the famine is not among the best actions that John can perform. The ranking is the same for the realist; which actions are ranked is different.

Eliteness has a role to play not only in settling which ranking speakers are using, but also in fixing a modal base. There are a number of sets of actions that could be ranked. For a speaker using a practical term without the ability role, the modal base is the set of actions that someone without any contingent limitations could perform in the agent's circumstance. There are other candidate modal bases. For instance, there is the set of actions an agent without limits could perform that do not require raising her

left pinky finger. There are many other gerrymandered sets of actions in the vicinity as well.

Speakers might be mistaken about the modal base at issue. Take a speaker who uses her moral term 'ought' without the ability role while at the same time harboring bizarre views about what is metaphysically possible, holding that it is impossible for there to be more than a certain amount of goodness in the world. Any amount of goodness beyond a particular threshold is, according to this speaker, impossible. Such a speaker is still capable of having a substantive disagreement with other speakers who use the moral 'ought' without the ability role. For instance take a dispute constituted by the following assertions

E : Jamie ought to end the famine;

F : It is not the case that Jamie ought to end the famine.

This is a substantive dispute in the case at hand. Both E and F are using 'ought' without the ability role. It is a fact that eliminating a famine is metaphysically possible for some agent, and would make the world better. F is mistaken about this fact (because she thinks that eliminating the famine would exceed the threshold for metaphysically possible goodness) and thinks she uses her 'ought' with a modal base that does not include the act o ending the famine. She is nonetheless having a substantive disagreement about it with E. But her use of 'ought' in accordance with her bizarre metaphysical views means that her use of the term alone is not sufficient to explain the substantive disagreement. We need external conditions on what settles the modal base beyond the speaker's intentions on the contextualist view; a realist can feel free to add a preference for eliteness to the list.

The contextualist approach to the semantics of practical terms makes it tempting to account for moral and normative disagreement in non-standard ways. On some views, speakers disagree over which context, which sets the ordering source and modal base parameters, to be in.[154] Other reject a descriptivist meta-semantics.[155] I have not argued against these views here. Instead I have simply sketched how a straightforward realist view is not *required* to adopt these departures. A realist who aims to treat practical claims as describing a privileged part of reality—the part that bears on what we should do—can use reference magnetism to explain the facts about disagreement in these terms. This remains an open possibility, even if the semantics of these terms as given by the contextualist view.

Accounting for substantive disagreements with practical language relies, for the realist, on a number of assumptions. I have adopted the primitivist view of eliteness, which holds that the facts about which properties are elite are not grounded in any further facts. Accounting for which disputes involving practical terms are substantive disagreements, and which are merely verbal, requires some assumptions about where these elite properties are located. I have argued that a view on which there are multiple highly elite moral and normative properties, including not only obligation simpliciter, but also the best state property, the property of being boundedly optimal, and the like, can account for this.

This gives a natural, realist-friendly explanation for some puzzling features of practical language. It rests on the realist's distinctive metaphysical claims about properties like obligation and moral rightness, in order to explain why it is that practical terms

[154] Silk (2016)
[155] Chrisman (2015)

are highly stable, and give rise to extensive disagreements between a wide variety of possible linguistic communities. Moreover, it can do so without overgeneralizing to the implausible thesis that every possible community that has a moral or normative term will be capable of entering into such disagreements. This is the **Universal Disagreement** thesis, which is false. The realist, on the picture I have sketched here, can explain why, simply by adopting plausible assumptions about the distribution of elite properties.

By itself, this would be a strong case for adopting a realism that is committed to the eliteness of moral and normative properties, and a meta-semantics that includes **Magnetism**. In the closing chapters I will extend the credentials of the view further. The additional benefits of the view start with an epistemological claim about what it takes to *know* claims about eliteness. With these claims in hand, the realist view can deliver further explanations of central claims, and also reply to more objections.

Chapter 5.

Knowledge, eliteness, and alternative theories

This book began with a characterization of which possible communities in modal space have practical disagreements with each other. There are all kinds of possible communities of language-users, and among those that use practical language, we can ask which of these communities genuinely disagree about moral and normative matters, and which of these communities are talking past one another, or are having merely verbal disputes.

The **Universal Disagreement** thesis holds that two possible communities must be capable of genuinely disagreeing, if they both use their 'ought' with the same practical role. For example, the fact that the communities are by stipulation both using normative terms is sufficient for them to be capable of a genuine disagreement.

While this is a tempting characterization of the extent of possible disagreements between users of practical language, it is too quick. As Chapter 2 argued, we need to know more about the possible communities in question before we diagnose them as disagreeing. It does not merely follow from the fact that they use 'ought' as a normative term alone, that they are referring to the same thing, and thereby disagreeing with each other. Similarly it does not follow that communities are capable of genuine disagreements if they use 'right' with a moral role. There are some possible communities who use 'ought' as a normative term, but are not speaking about the same thing we are. The **Universal Disagreement** thesis is false.

This does not, however, mean that the **Universal Disagreement** thesis is completely unmotivated. Instead, while it is false, it is a natural generalization off of some insights about the extent to which practical terms are stable. Across modal space, there are

many different uses of 'ought' as a normative term that do have the same referent. What is not supported is the inference to the claim that *all* possible users of normative language are referring to the same thing. What needs to be explained, then, is not the **Universal Disagreement** thesis, but rather something more limited: while *many* possible communities have normative disagreements with each other, not all of them do. The extent of normative disagreements across modal space is pervasive. But there are limits to the extent of practical disagreements. Thus, a weaker thesis—which I call **Robust Disagreement**—is all that we should accept.

How should we explain **Robust Disagreement**? It is not the mere existence of a shared practical role that explains why communities who use their practical terms differently can have disagreements with each other. This would explain too much, as it would entail the **Universal Disagreement** thesis. Chapters 3 and 4 outline an alternative explanation, which starts from the realist assumption that moral and normative properties, such as the properties of moral rightness and normative obligation, are a part of reality itself. This is, of course, just a slogan, and so the first task is to explain how the realist should understand it. The realist can use a metaphysical distinction between those properties that are elite, and those that are not, to do this. What the version of realism I am developing holds is that moral and normative properties exist as an elite part of reality. They are not unique in this, as there are non-moral and non-normative properties that are also elite.

Eliteness itself is a primitive feature of certain properties and not others. We can distinguish the elite from the non-elite by the roles that they play: the elite properties show up in the law-like generalizations of sciences and other theoretically sound disciplines; claims involving elite properties are projectable, and elite properties confer

some measure of similarity on those things that share them. This is the eliteness-role. I have defended in Chapters 3 and 4 the claim that elite properties are reference magnets, and that this provides a natural way for the realist to explain the **Robust Disagreement** thesis.

The two main components of this explanation are as follows. First, moral and normative terms are highly stable—and so disagreement is pervasive across possible uses of practical language—because there are elite properties of moral rightness and normative obligation, which serve as reference magnet for typical uses of the moral 'right' and normative 'ought'. Since there are no other properties that are both similar to rightness and obligation, and highly elite, these properties are *unique* reference magnets for many uses of practical vocabulary. But, second, these are not the only elite properties that fit *every* possible use of practical vocabulary. There are others as well, although (given the uniqueness assumption) they are not very similar to obligation. These properties fit uses of practical terms, but only once they are used with roles certain kinds of additional roles. Reference to these alternative moral or normative reference magnets is, while possible, not a nearby possibility in modal space for speakers who do not use their practical terms with the relevant additional roles.

This explanation of the **Robust Disagreement** thesis (and the failure of the **Universal Disagreement** thesis) leaves a number of questions unanswered. As I have emphasized, the eliteness of a property—and so, in particular, the eliteness of the normative property of obligation—is a primitive, unexplained fact. The explanation of **Robust Disagreement** assumes, without an explanation, that the elite normative properties are distributed in was that make the reference-magnetic explanation work. This assumption, given the primitivist approach to eliteness, is in some sense justified.

Primitivism predicts that we cannot give a further constitutive account of what makes these properties elite rather than others. And the success in explaining a puzzling explanandum counts somewhat in favor of the picture.

This raises the question of whether there is anything we can say in favor of the claim that obligation is elite, and that there are other normative properties that are elite, but not too similar to obligation. I will provide a partial answer to this question. The details are below, but the general picture is this: first, without giving an account of what *constitutes* the eliteness of obligation (which would be incompatible with primitivism), we can still give an account of how we can *know* that obligation is elite. This epistemological claim entails that obligation is elite, owing to the factivity of knowledge. Or, more accurately, I will be arguing for a possibility claim: that we can have the relevant knowledge. While actual knowledge is factive, possible knowledge is not by itself factive; some false propositions are possibly true, and if they were true, then we would know them. We need an extra premise: that the eliteness-facts are necessary, if true. It follows from this assumption, plus the factivity of knowledge, that if we can know that such-and-such property is elite, then it is elite in the actual world.

Second, the explanation trades on a basic intuition about *which* morally and normatively relevant properties are candidate referents for different possible uses of practical terms. This is the intuition that a possible community uses a practical term to refer to something other than moral rightness or normative obligation when their use characterizes a distinct, theoretically interesting property that is relevant to moral and normative theorizing. Thus there is a connection between the generalizations of different areas of theorizing about morality and normativity that we engage in, and the possible communities that use their practical language to refer to properties that are

distinct from rightness and obligation.

To make and develop this explanation, I will in §1 outline the connection between the law-like generalizations of our theories and what we can know about which properties are elite. This is an elaboration of one component of the eliteness-role, coupled with the non-skeptical assumption that knowledge of the law-like generalizations about various subject-matters is possible. Then I will fill out this picture sketching some structural features of knowledge in §2. This relies on the connection between knowledge and the absence of *risk* of a false belief. The relationship between knowledge and risk will serve not only to illustrate the specific sense in which we can have knowledge of which properties are elite, but also why such knowledge is not guaranteed, or necessary. In §3 I will apply these results to the central explanandum of this book by deriving the conclusion that we can know that certain properties from different areas of moral or normative theorizing are elite. The elite properties that fit possible uses of practical language are not limited to moral rightness and normative obligation. Instead, or practice of ethical theorizing should lead us to expect, given the theses of SS1-2 that there are elite properties that roughly fit the uses of communities in cases where **Universal Stability** fails. In §4 I present a generalization of this explanation.

5.1 Knowledge and laws

The primitivist holds that facts about which properties are elite are primitive. They are not defined or grounded in further facts. That is: it is a fact that the property of being charged is elite, and it is a fact that the property of being a living organism is elite. If the realist view elaborated here is correct, it is a fact that the normative property

of obligation is elite. The primitivist view is motivated by general considerations: the proposed definition in Lewis (1983) of eliteness for non-physical properties does not provide a notion of eliteness that serve can undergird reference magnetism, or play other components of the eliteness-role.[156] Lewis himself endorsed something like primitivism for what he called the "perfectly natural" properties, which he took to be the elite properties from fundamental physics, so the idea is not incoherent. The view of eliteness I am working with here simply extends the primitivist treatment to the elite (or elite-to-some-degree) properties that are discovered by other disciplines including, crucially, the various areas of ethics.

The primitivist is not at a loss for how to identify the elite properties; Lewis held that we can know which properties are perfectly elite by looking to fully developed physics:

> Thus the business of physics is not just to discover laws and causal explanations. In putting forward as comprehensive theories that recognize only a limited range of natural properties, physics proposes inventories of the natural properties instantiated in our world. (Lewis, 1984, 364)

A more generous account of which properties are primitively elite will also need to be more generous with the methods for identifying the elite properties. This is consequence of the more general claim that the elite properties are nut just physical properties, but include any property that plays the eliteness-role. The claim that a property P plays the eliteness-role amounts to the following: P confers similarity on its bearers, P features in projectable generalizations, and law-like generalizations include P. When physics discovers laws and causal explanations, it proposes a set of law-like generalizations that include terms like 'mass', 'charge', and so on—that is, it claims that the

[156] See Williams (2007) and Hawthorne (2007) and the discussion in Chapter 3.

referents of 'mass' and 'charge' play one component of the eliteness-role. This section will unpack and extend this idea.

5.1.1 Laws and theories

First it is worth clarifying what Lewis's approach to the relationship between the laws of physics and the perfectly elite properties is. Here is an instructive quote from Dorr and Hawthorne (2013):

> The claim is not, of course, that every word that physicists use is to be counted as expressing a perfectly natural property: Lewis would not be sympathetic to the suggestion that *being a Nobel Prize winner* is perfectly natural. Even if we only looked at the words the physicists use when stating what they call 'laws', we will be apt to find our list of perfectly natural properties contaminated by properties like *being a measurement* [...] (Dorr and Hawthorne, 2013, 18-19)

The first point to take from this is that the relationship between actual theorizing in physics, and the elite physical properties, is not simple. I do not have a theory of what the relationship between physicists' theorizing and the eliteness of mass and charge is, and will instead simply assume that there is some important connection between the fact that mass is elite (and can be known to be so), and the fact that mass features in physical theories. What the relevant kind of "featuring in" is that distinguishes mass from Nobel Prize winners and measurements is not something I have more to say about.

A second point is that this is a claim about the relationship between a particular theory (physics) and its law-like generalizations. I will distinguish these from each other. The theory is an entity that physicists create and develop; they test the theory, and refine it. Physical theory has changed significantly over the centuries. Laws, and law-like generalizations, by contrast, are real-world constituents that exist regardless

of whether any scientists bother to formulate, test, or believe them. While physical theory has changed, the laws of physics have not.[157] The epistemological claim I am advancing, following Lewis, is that by knowing that a particular theory is true, we can know the relevant law-like generalizations, and thereby know the eliteness-facts about mass, charge, etc.

Third, this picture can be generalized beyond physics, and what I say in the following explicitly relies on this generalization. Biological theories uncover law-like generalizations about organisms. Chemistry discovers law-like generalizations about hydrogen. It is clear how, in rough outline, we should continue the pattern: the generalizations uncovered by geology, neurology, psychology, astronomy, aim at discovering law-like generalizations. The properties that feature in these laws, such as cells, acids, neurons, etc., are also elite.

Fourth, the extension does not have to stop at theories that fall under the heading of the "sciences" in contemporary parlance. So far we have been focusing on scientific theories that are broadly empirical in nature, such physics, biology, chemistry, and the like.[158] But in principle there is no reason to restrict our attention to theories from the empirical sciences. There are theoretically legitimate disciplines that discover relevantly similar laws in mathematics and logic. We should not rule out that these disciplines relate to laws in an analogous way. I will assume this, and add an assumption that will be very natural for the realist: that ethics, and normative theorizing more generally, has the same epistemological structure. I have already outlined the assumption, in Chapter 3, that ethical and normative properties have the same metaphysical

[157] I assume that versions of both Humeanism (Lewis, 1994) and necessitarianism (Armstrong, 1983) are consistent with what I say about fundamental physical laws. I will also include discussion of "higher-level" or "special science" laws below, and will not advance a theory of the metaphysics of these laws here. Rather I simply assume that a fully general account of lawhood is available here.

[158] Dorr and Hawthorne (2013, 18) call the Lewisian view of the epistemology of eliteness "Empiricism".

status as physical, biological, chemical (and so on) properties, as they share the metaphysical status of being elite. Now I am adding the assumption that ethics relates to lawhood and theorizing in the same way as well: good ethical and normative theories discover law-like generalizations about moral rightness, normative obligation, and so on.

Finally, it is worth point out that there is a kind of benign circularity in the primitivist view of eliteness, when combined the assumption that elite properties appear in the (appropriately developed) generalizations of our best theories. The elite properties, which are reference magnets, can be known to be such in virtue of their appearance in law-like generalizations. But what makes them appear in law-like generalizations rather than other, nearby gerrymandered properties? And why do our theories refer to them? It is, in part, their eliteness.[159] Sentences stating the laws of chemistry refer to hydrogen and not hydrogen-before-3000AD. Here the explanation is that hydrogen is elite, and hydrogen-before-3000AD is not. So eliteness explains why hydrogen shows up in the chemical laws, which in turn explains how we can know that hydrogen is elite, which then explains how we can know that it is hydrogen (rather than a nearby gerrymander, such as hydrogen-before-3000AD) that shows up in the statement of a law. This would be a vicious type of circularity if, contra the primitivist, the the content of law-like generalizations metaphysically grounds eliteness-facts. It cannot be that (i) the eliteness of hydrogen grounds (in part) the fact that hydrogen shows up in chemical laws, and also (ii) the fact that hydrogen shows up in the chemical laws grounds the fact that hydrogen is elite. The primitivist denies (ii). The laws allow us to *know* that hydrogen is elite, but do not ground this fact.

[159] See also Hawthorne (1994)

This idea that our theorizing, when done well, provides insight into the metaphysical status of the entities and properties it invokes is not an original idea. This is Sider's epistemology of "Structure" in Sider (2012):

> A good theory isn't merely likely to be true. Its ideology is also likely to carve at the joints. For the conceptual decisions made in adopting that theory—and not just the theory's ontology—were vindicated; those conceptual decisions also took part in a theoretical success, and also inherit a borrowed luster. So we can add to the Quinean advice: regard the ideology of your best theory as carving at the joints. We have defeasible reason to believe that the conceptual decisions of successful theories correspond to something real: reality's structure. (Sider, 2012, 12)

The only innovation here is the proposal that we treat ethical theorizing, broadly construed, similarly. To some extent, this is not surprising. A realist approach to morality and normativity takes as its starting point that moral and normative properties are features of reality, in the same way that other real features of the world are. We think that science and other broadly empirical disciplines are guides to some parts of reality; a realist about morality should then think that ethical theorizing does the same. Since I am working with the view that the objective parts of reality are metaphysically elite, the connection between ethical theorizing and knowledge elite ethical properties is a very natural one to make.

5.2 Knowledge and epistemic risk

Suppose, as I have outlined above, that the law-like generalizations of certain theoretical disciplines can give us knowledge of which properties are elite. It is worth adding some detail to what this claim about knowledge amounts to.[160] This episte-

[160]One important question, which I will not address in what follows, is the need to distinguish between "legitimate" theoretical disciplines that produce genuine law-like generalizations, and pseudo-theorizing that does not. Alchemy, astrology, and phrenology present theories that correspond with observed data, at least to some extent. Utilitarian ethical theorizing can be made to correspond with ethical truths, to some extent. This does not mean that these theories capture laws, and provide the resources to know that

mology of eliteness is claim about what it is *possible* to know, twice over. It says that it is possible to know the contents of good theories. And, it says that it is possible that, for someone who does know the relevant theory, that they know about the eliteness of the properties it references.

I will develop this idea with simple necessary condition on knowledge. This is an *anti-risk* principle: roughly, a belief is knowledge only if it is not at risk of being false. This idea needs refinement, but the basic idea is that if realism about morality and normativity is plausible, it then among its consequences must be the epistemic fact that moral and normative beliefs are not at risk of being false, in the relevant sense.[161]

The relevant notion of *risk* is to be cashed out in terms of what goes on in nearby worlds, or worlds that could easily have obtained.[162] When one risks dropping one's phone in a pool by standing on the edge of the pool and tossing the phone in the air, this amounts to the existence of a nearby world where one tosses the phone in the air and it falls into the pool. Likewise when a belief is at risk of being false, this amounts to the existence of a nearby world in which the belief is false.

I will say that, when a belief could easily be false in the relevant sense, it is subject to *epistemic* risk. Beliefs that are subject to epistemic risk are subject to a kind of risk that is incompatible with knowledge. Calling the risk "epistemic" serves to distinguish it from other kinds of risk that may not—or are not by definition—incompatible with knowledge.[163]

---

the properties referenced by these "laws" are elite. Knowledge of which properties are elite comes only from genuine theories that uncover genuine laws. It may be possible to give an informative and general characterization of the difference between the disciplines that are capable of stating laws, and those that are not, although this is not obvious. See Lakatos (1974) and Hansson (2014).

[161]Similar points will apply to epistemic justification. One way to lose justification for a belief is to learn that it could easily have been false in a sense that is incompatible with knowledge. So if knowledge is absent because of the presence of a kind of objectionable risk, one will lose justification for a normative belief when one learns that the relevant kind of risk is present.

[162]*Cf.* "safety" principles in Sosa (1999), Williamson (2000), and Pritchard (2004).

[163]I do not intend to commit, as some have, to the claim that epistemic risk constitutes the best analysis or

5.2.1 *Risk, similarity, and belief-forming methods*

At a first pass, a belief that is subject to epistemic risk is one that could easily have been false—that is, there is a nearby world where the belief is false. But this is just a first pass, and several refinements are needed. First, a belief is at risk in this sense only if there are *similar* beliefs that are false in nearby worlds. I can know that I had breakfast this morning even if there is a nearby world where I misremember the name of a new acquaintance. A belief about someone's name is not similar to a belief about what I had for breakfast, and if one is false in a nearby world, the other is not at risk of being false.

But it will not do to restrict these knowledge-destroying false beliefs to beliefs that are identical in content, either. If one is guessing at the answer to questions about the sums of moderately large numbers, then one's correct guesses won't have nearby worlds where the same belief is false. If one correctly guesses that 634 + 399 = 1033, then one has a true belief, and moreover this very belief is not false in any nearby world (in all nearby worlds, 634 + 399 = 1033). Correctly guessing does not, however, bring knowledge. If one is guessing at the relevant sums, then even if one actually gets the answer right, there is a nearby world where one instead comes to believe a related but false claim—for instance that that 634 + 399 = 893. This belief is sufficiently similar to one's actual belief, and since there are nearby worlds where one has false beliefs like this when one is guessing, one's actual true belief that 634 + 399 = 1033 is at risk and is not knowledge.

Here is a second qualification: not all similar false beliefs are incompatible with

informative characterization of knowledge. So long as it is plausible that the presence of epistemic risk is a reliable indicator of the absence of knowledge (even if it doesn't constitute such an absence), the notion will be useful for assessing the relationship between the knowledge constraint and other issues for the realist view.

knowledge in this way. Some nearby similar false beliefs are arrived at in a suitably different way, and so do not put one's actual beliefs at risk, in the relevant sense. If I happen to see Gabe walk past my office, I know that Gabe is on campus (we can suppose this is true even if I have no other evidence that Gabe is on campus and would otherwise have believed he is somewhere else). There is a nearby world where I don't look up the minute Gabe walks by my office, and hence continue to believe that he is not on campus. But my true belief that Gabe is on campus isn't at risk just because there is a nearby world where I don't look up, and on that basis have a false belief.[164] The reason is that the beliefs are formed by very different processes, as the causal processes that produce each belief are very dissimilar. One involves perception and the other involves an inference on the basis of by knowledge of Gabe's usual whereabouts. Thus if a belief is at epistemic risk, there must be nearby false beliefs that are both similar in content *and* similar in respect of the causal processes that produce them.

This is not a comprehensive or definitive account of the conditions on knowledge.[165]. But it provides us with a range of structural features that accompany knowledge, and which allow the primitivist to develop realism in a number of ways.

5.2.2 Eliteness and contingent knowledge

Given an anti-risk condition on knowledge, it follows from the fact that we can know the law-like generalizations of our theories, that we can believe these generalizations without being at risk of believing a false similar generalization, in the sense

[164] Cf. Pritchard (2004)

[165] Some, including Hiller and Neta (2007) and Setiya (2012), argue that the conditions here are not perfect proxies for knowledge. I will not engage with the criticism here since my claim is not that knowledge *requires* satisfaction of these conditions, but only that it is typically accompanied by them. I provide some additional defenses of conditions like these in Dunaway (2017b) and Dunaway (2018)

outlined above. That our theories can produce knowledge is not a claim that I will defend here, since any non-skeptic should grant it. Instead I will simply outline some structural features that accompany this fact, before moving on to knowledge of which properties are elite.

One important fact to note is that I am only claiming that it is possible that we know the relevant law-like generalizations; not that we *do* know them. That we possibly know these relevant law-like generalizations is a claim about modal space: there is a possible world where we truly believe them, and moreover this world is one where beliefs in these generalizations are not subject to epistemic risk. This is consistent with the actual world being one where our epistemic situation with respect to the laws is fairly impoverished, fo a number of reasons. (i) It could be that in the actual world our theories are underdeveloped. (ii) It could be that our theories have been developed in a misleading direction—they include false assumptions that will make uncovering the actual laws with further developments impossible. (iii) It could be that our theorizing is akin to what goes on in non-genuine, pseudo-scientific theories. Even if one of (i)-(iii) describes our actual situation, the actual world isn't a world that is relevantly close to every other possible world. We *could* be in a world where we have true beliefs that are free from risk, even if our actual situation is impoverished.

Given this assumption about modal space, it is possible to know the law-like generalizations about a subject-matter. What is more important for our purposes is knowledge of facts of the form *P is an elite property*, where the knowledge is arrived at on the basis of knowledge that *P* features in the law-like generalizations discovered by a good theory. The same points about the contingency of this knowledge apply here: while it is possible to truly believe that *P* is elite, and for that belief to be free from risk, it is

not necessarily the case that our actual beliefs about which properties are elite meets this condition.

One simple way for this to occur is when one is not in a position to know the relevant law-like generalizations. Suppose one is at risk of having a false belief when believing the law-like generalization

G All *F*s are *G*s.

One might, on the basis of the fact that *F*s are mentioned by the generalization **G**, infer that *F*-ness is elite. But given that one is subject to epistemic risk when believing the generalization, one is also at risk in believing that *F*s are elite.[166] So, epistemic risk in beliefs about the contents of one's best theory will also put one's beliefs about which properties are elite at risk.

This schematic picture points to a second way in which one can fail to know eliteness-claims *even if* one is in a position to know the law-like generalizations of a theory. This is because knowing the fact that *F*-ness is elite *also* requires knowing, in normal circumstances, conditionals of the form

E If all *F*s are *G*s is a law-like generalization of our best overall theory, then *F*s are elite.

Clearly, one can know *what* one's best theory says, without knowing that the properties that feature in it are elite. But there are some ways of failing to know the second claim, which put our beliefs about eliteness-facts at risk, even if our beliefs about the relevant

[166]More concretely: suppose one's belief in the generalization expressed by 'all *F*s are *G*s' is at risk because one could easily have believed a false theory, which contains the (false) generalization 'all $F*$s are *G*' instead. In the world where once has this false belief, one will, by virtue of having the disposition to infer eliteness-facts from the generalizations of one's best theory, also have the false belief in claim about eliteness that is expressed by '$F*$-ness is elite'. This is a false belief in a nearby world, if the belief in the false theory is. It is also similar (both are beliefs about eliteness), and formed by the same method.

generalizations are not subject to epistemic risk. Here are several ways for this to happen. Importantly, it need not be that one actually meet any of the conditions specified below; it is enough that one could easily have met them, for knowledge of **E** to be absent.

First, one could believe that too few theoretical disciplines involve law-like generalizations that reference elite properties. For instance, one could hold Lewis's view that only the laws of physics involve elite properties. Second, one could believe that too many theoretical disciplines involve law-like generalizations that reference elite properties. For instance, one could believe that the property of being a Virgo is elite, because it features in the generalizations of astrology. Third, one could believe that law-like generalizations from our best overall theory have no relationship to claims about eliteness at all—for instance, one might think that what is elite is purely a question for a priori metaphysics. Moreover there are ways of filling in each scenario so that the content and process by which the erroneous beliefs are formed are sufficiently similar to give rise to epistemic risk.

These points show that knowledge of the eliteness-facts is, on the view developed here, not guaranteed simply by knowledge of the law-like generalizations of our theories from physics, chemistry, biology, and (as I am claiming here) ethics. Simply adopting a non-skeptical view with respect to our knowledge of the content of well-developed versions of theories of each kind—that is, knowledge of claims in the form of **G**—is not enough to guarantee that we can know the eliteness-facts on the present view. We need in addition to be in a position to reliably reason, in nearby worlds, using **E**, to true claims about eliteness.

This is just a sketch of a general epistemology of eliteness. In the next section I

will develop it, with application to our possible knowledge of moral and normative theories, and in particular, knowledge of which moral and normative properties are elite. The upshot will be that, given the epistemology of eliteness sketched here, we should expect the realist view I have developed to hold that there are elite moral and normative properties which are distributed in a way that makes the reference-magnetic explanation of disagreement that I developed in Chapter 4 work.

5.3 Ethical theories and elite properties

Chapter 2 presents two broad features of moral and normative terms that need to be explained. The first is the striking range of stability of these terms: as Horgan and Timmons and others have emphasized, there is a wide range of communities throughout modal space who use their practical terms differently, but still manage to be talking about the same thing. This is the phenomenon captured by the Pervasive Disagreement thesis. But, as I have emphasized, this does not generalize to the **Universal Disagreement** thesis. There are some possible communities who use 'ought' as a normative term, but fail to talk about the same thing that we refer to with our normative 'ought'. **Universal Disagreement** fails because there are some possible communities who succeed in using practical terms to talk about different things.

Chapter 4 has already laid out the form of the realist explanation of this phenomenon that I am exploring. The goal of this chapter is to use the epistemology of eliteness outlined above to add something to this explanation. This is the claim that we can *know* that the elite properties are distributed roughly as the Chapter 4 explanation needs them to be.

5.4 Ethics and alternative practical subjects

I assume there is some true moral theory. Our current state of ethical theorizing may not be very close to such a theory, but, I assume, such a theory exists. Even if we don't know the theory, and even if we wouldn't come to know the theory by improving our current theorizing, the theory is known by some possible people. The properties mentioned in the law-like generalizations of this theory are elite. Below I sketch what this idea adds to the realist theory that has eliteness and reference magnetism at its core. First, I fill out the explanation of why moral and normative terms are highly stable, in a way that makes the **Robust Disagreement** thesis true. Then, I turn to the explanation of why **Universal Disagreement** is false.

5.4.1 Shared reference

Recall the difference between the communities in a simple version of a Moral Twin Earth case as Horgan and Timmons describe it. There are two communities, and they differ in their use of a practical term:

> 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 [...]
>
> [O]n Moral Twin Earth, terms in people's uses 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 [...] But these are non-consequentialist moral properties [...][167]

It will be very natural for each community to develop different theories about the property their moral terms refer to. One will develop a kind of deontological the-

[167] Horgan and Timmons (1992b, 245)

ory; plausibly the "regulation" of their use by a particular property will, over time, cause them to theorize as if the property that is moral rightness is the deontological property. (If the relevant version of deontology has the avoidance of rights-violations at its core, then they will accept after sufficient theoretical reflection a generalization along the lines of 'avoiding violating rights is obligatory'.) For similar reasons, the consequentialist community will do the same thing in their theorizing, except with the relevant consequntialist property featuring in the law-like generalizations of their theory. (The Utilitarian version of the consequentialist theory would include the generalization 'happiness-maximization is obligatory'.)

At most one of these communities will accept the true theory of moral rightness. Perhaps neither does: the theories that develop in these communities are not exhaustive. But there is a true theory, and even if neither of the communities in the Moral Twin Earth case comes to believe it, it is still possible that some community does. Such a community can know the theory, and on this basis know that the property mentioned in this theory is the elite property of moral rightness. This amounts to the claim that some possible community can believe the generalizations of this theory, which reference the relevant elite property, without being subject to epistemic risk. I defend this claim in light of the fact that at least one of the communities in the Moral Twin Earth case is mistaken in the next chapter.

Since some community can know that the property which is moral rightness is elite, it follows that this property *is* elite.[168] So there is an elite property of moral rightness in the vicinity of the use of 'right' by the communities in the Moral Twin Earth scenario. Then we can develop an explanation of why they are having a substantive

[168]This inference relies on the factivity of knowledge, plus the assumption that if P is elite, then it is necessarily elite.

disagreement based on **Magnetism**: since both of the Moral Twin Earth communities use their 'right' as a moral term, their use shares something in common: each community uses the 'right' with the same moral role. Moral rightness fits this role perfectly. Moreover, it is elite. So, it scores highly on the two factors that matter for reference-determination: fit and eliteness.

For at least one of these communities, other aspects of their usage will fit poorly with obligation. If the consequentialists are wrong, they will apply their moral 'right' to happiness-maximizing actions that are not obligatory. A meta-semantic theory that includes reference magnetism will reject the view that reference requires perfect fit. Since even the consequentialists will use their moral 'right', by using it with a moral role, they use it in a way that fits very well with the property of moral rightness, and so the fit is not awful. The eliteness of moral rightness then tilts the balance in favor of determinate reference to this property, even though the community in question has many false beliefs about what rightness is.

It is worth emphasizing three ways in which the epistemology of eliteness adds to this solution. First, it explains the negative verdict: that there is not an additional reference magnet that the consequentialist community refers to. Or, more generally, it explains why it is not the case that there are two distinct properties that each fair equally well on the use-plus-eliteness metric. At least one community is using their moral 'right' in accordance with a false moral theory. The properties that feature in the generalizations of this theory are not elite. Second, no part of this solution requires that we actually know which property is the elite property of moral rightness. As long as there is some true moral theory that can be known—regardless of whether we actually know what it is—there is guaranteed to be some elite moral property. The pretense

that we know that one of the Moral Twin Earth communities is right makes exposition simpler, but is no part of the official explanation. Third, there is no circularity in this account. The fact that there is a moral theory that can be known is not what grounds or constitutes the eliteness of the relevant moral property. The eliteness of this property is primitive; moreover what makes it the case that the statement of the correct theory refers to obligation is, in part, the fact that it is elite. What our (possible) knowledge of this theory adds is the grounds for assuming that there is one, and not more than one, elite property that fits reasonably well with the usage of the moral 'right' across a wide range of modal space.

Since the two communities in the original Moral Twin Earth case are not the only communities that use their moral 'right' with a moral role, this point will generalize to other possible communities. Moreover, I assume that similar assumptions hold for the normative 'ought', since there is a distinct true theory of all-things-considered normative obligation, whose law-like generalizations can be known as well. If so, we have a sketch of an explanation for why we should expect **Robust Disagreement** to be true, on the realist theory developed here.

Below I will explain why we can also expect there to be other reference magnets that prevent this explanation from overgeneralizing. While we should expect that there are unique and highly elite moral and normative properties that support an explanation of **Robust Disagreement**, the same method does not imply that these are the only elite properties that are morally or normatively relevant. So reference magnetism does not support **Universal Disagreement**.

5.4.2 *Alternative theories*

At a first pass, the reason is that there are theories which are relevant to morality and normative theory, but are distinct from the true theories of moral rightness and normative obligation. They differ from the true theories not because they are false, but because they concern a different subject-matter. Their subject-matter still belongs to ethics, broadly construed. But ethics, broadly construed, concerns more than what is morally right, and what we all-things-considered should do. These are (perhaps the) central topics for ethics, but not the only ones. There are other subjects that are ethically relevant, and we can develop theories about these subjects as well. Moreover these theories can be true, although they are true of something other than rightness or obligation. Thus they are not distinguished from the true moral and normative theories by being false theories of morality and normativity; they are *potentially true* theories of some other subject-matter. I will call these *alternative theories*.

Here is one example of an alternative theory, in this sense. What is sometimes called the "evaluative" ought in English occurs in assertions such as

There ought to be world peace.[169]

The truth-conditions for an evaluative use of 'ought' in contexts like this are roughly as follows: 'ought' prefaced to a term that designates a state of affairs generates a true sentence just in case it would be best for the relevant state of affairs to obtain. These are facts that are unrelated to what any particular agent is able to do: it would be best if there is world peace, even if no one can bring world peace about.[170]

[169]There is a linguistic issue that should be separated from a metaphysical issue here: in English, 'ought' plausibly can express either an evaluative, or what I am calling a normative or ethical notion. There may be linguistic distinctions between evaluative and normative uses of 'ought' (cf. Schroeder (2011)) but not the main issue for the present point.

[170]Contra Geach (1982). See Ross (2010).

What constitutes the state that it would be best for the world to be in is a theoretical question for ethicists, although one that is distinct from the question of what we morally ought to do. Nonetheless there is, as in the case of theories about moral rightness, a true theory about the subject-matter; non-skeptics will maintain that we can know it. The epistemology of eliteness I have sketched then entails that there is another elite property: it is whatever property is identical to *being the best state for the world to be in*. This is the "best state" property from Chapter 2. Since the best state property features in the law-like generalizations of theoretical investigations into the nature of evaluative properties, we can know that it is elite. Such a theory is an alternative theory to a theory of moral rightness; the best state property is distinct from moral rightness.

This provides another reference magnet for practical vocabulary. In the normal case (as in English) the words we use to pick out the evaluative notion of the best state for the world to be in, we do so by using vocabulary that does not have a normative or moral role attached to it.[171] However while this is the actual way we use our practical terms, there may be other possible communities which, for some reason, do not use their terms in this way. There *could* be a community that uses a practical term 'ought' that fits extremely well with whatever elite property constitutes the best state property. At some level there is something confused about this—the nature of the concept would make it extremely common for speakers to judge that they are making conceptual mistakes. A role that characterizes the best state property will not characterize properties that play a normative role. So, to some extent a community that uses their practical terms in this way does not have a term that can be perfectly fit

[171] Though see Tappolet (2014).

by any one property.

But reference-determination does not require a perfect degree of fit with use. One reason why **Universal Disagreement** fails is that practical roles do not automatically outweigh other roles, plus the eliteness of the best state property. It fits what we called the "ability role" in Chapter 2 fairly well, but does not perfectly fit the moral or normative role that constitutes the distinctive feature of a practical term. However this latter feature is not automatically disqualifying: reference-determination does not require a perfect fit with every aspect of use. And since the best state property is elite, it is then plausible that the best state property does the best job of maximizing fit with use plus eliteness. It is the referent of some possible uses of practical language.

This is a consequence of the existence of alternative theories, plus the machinery I have introduced here. We can know the generalizations of these alternative theories; so, we can know that there are other properties, aside from moral rightness and all-things-considered obligation, that are elite. Since these properties can fare well on the fit dimension of reference-determination of some possible uses of practical language, they allow for counterexamples to **Universal Disagreement**.

This is the schematic explanation of why the Ability Twins from Chapter 2 do not disagree with speakers who use practical terms with the meaning that they have in English. A similar point applies to the community that uses a practical term with the bounded optimality role. The core points are familiar: first, there is a legitimate area of theoretical investigation that develops the principles of bounded rationality. (Perhaps this area is underdeveloped.) Second, there is a correct version of this theory, and its law-like generalizations can be known. And third, the properties that feature in these generalizations can on this basis be known to be elite.

So, there is an elite property of bounded obligation. As before, I am not assuming any particular version of this theory, so I will not take a stance on which property bounded obligation is. Nonetheless, it is a safe assumption that it is not identical to obligation. There is another reference magnet, aside from obligation, and this property will be a decent but not perfect fit with the practical roles that are characteristic of practical terms. Possible speakers that also use their practical terms with the bounded optimality role may be such that the best overall fit for their use is bounded obligation. Since the property is elite, it is a reference magnet, and so other properties that are less good as fits with use will not be more eligible referents. Reference magnetism predicts that these communities will be counterexamples to the **Universal Disagreement** thesis.

Finally, we can repeat the point for the property of psychologically feasible obligation. A good theory of psychologically feasible obligation can be known, and on that basis a particular property which is psychologically feasible obligation can be known to be elite. This property will be a decent but not perfect fit with the practical roles that are characteristic of practical terms. Possible speakers that also use their practical terms with the psycholgoical feasibility role may be such that the best overall fit for their use is psychologically feasible obligation. Since the property is elite, it is a reference magnet, and so other properties that are less good as fits with use will not be more eligible referents. Reference magnetism also predicts that these communities will be counterexamples to the **Universal Disagreement** thesis.

5.5 A generalization: the Role-Theory Connection thesis

Thus far I have given a piecemeal explanation of the failure of practical terms to be stable to the extent required by the **Universal Disagreement** thesis. Given the

epistemology sketched here, some particular properties from some alternative theories are candidates for eliteness. However this should not give the impression that these are these are the only cases where stability for practical terms fails. In closing I will suggest a general thesis about the kinds of cases where we should expect stability for practical terms to fail.

The informal idea is that possible users of practical terms will not disagree when they are using their terms with roles which are characteristic of alternative practical theories. The examples in the previous section gave some examples of alternative theories: theories about the best state property, or bounded optimality, are not the same as theories of moral rightness or normative obligation. And the communities that are talking about these properties, even when using a practical term, are those that also use their practical term with roles that are characteristic of the best state property, or bounded optimality. In general, the hypothesis goes, whenever a community uses a practical term with a role that is characteristic of some alternative theory, then they are candidates for possible speakers who do not disagree with ordinary users of practical terms.

This idea can be refined and qualified in various ways. The result will be a set of generalizations that are mostly speculative. They capture an interesting and intuitive hypothesis about why practical 'ought's are robustly, but not universally, stable. And there is a natural explanation of why these generalizations are true, if elite properties are reference magnets. But these are not theses that I will argue for in detail here. Instead I simply present these generalizations as one general picture of what is going on in the first four chapters of this book.

Begin with the idea that practical terms are stable so long as they are used with the

same moral or normative role only. This is a plausible characterization of the **Robust Disagreement** thesis, and is the kind of thesis that the original Moral Twin Earth case supports. We can call it the **Role-Theory Connection** claim:

Role-Theory Connection Two communities use a practical term 'ought' to refer to the property P if: (i) there is a role r and each community uses 'ought' with role r, (ii) r approximates a true theory T, and (iii) T refers to P in its law-like generalizations.

Since, in the original Moral Twin Earth case, each community uses their term with a moral role, they satisfy component (i) of **Role-Theory Connection**. The moral role characterizes the property of moral rightness, so I will say that the role *approximates* the theory that characterizes moral rightness, in component (ii). The role is an aspect of use by a community: in the original Moral Twin Earth case, the use of each community includes the moral role, since their use is connected with feelings of blame and the like. The role approximates the true moral theory, rather than some other kind of theory (or no theory at all) because moral theory says that moral rightness is the property that makes failure to perform an action blameworthy, etc.[172] Since there is a specific property of moral rightness that this theory refers to, component (iii) is satisfied. **Role-Theory Connection** then says that the communities in question share a referent: 'right' is stable between them.

But this idea is not complete on its own. As we have seen, the fact that two communities share one aspect of use does not mean that they will necessarily be talking about the same thing. **Role-Theory Connection** glosses over this. We should expect to find communities who share a role r in virtue of their use pf a practical term but

[172]But the use only approximates the claims made by moral theories. Utilitarians will have views about when, and why, actions are blameworthy that depart from what the true moral theory says.

differ substantially and in the right ways in other respects, and thereby are talking about different things. The lesson from Chapters 1 and 2 is that a normative or moral role does not exhaust use of a practical term, and for some communities the additional aspects of their usage of their practical term overrides these roles.

Strictly speaking, then, stability between communities is only guaranteed when there is a *unique* connection between the role they use their 'ought' with, and the role outlined by a true theory of a practical subject-matter. That is, what is required is for shared reference is that there is only one theory which is approximated by the roles with which these communities use their 'ought'. Since the two communities in the original Moral Twin Earth scenario use their 'ought' with the role characterized by moral theory and no other roles that approximate other theories, they are guaranteed to be talking about the same thing, in spite of other differences in use between them. But the guarantee does not extend to communities that use their practical terms with additional roles, that approximate the properties characterized by different theories. A generalization of this idea is the **Unique Role-Theory Connection** thesis:

Unique Role-Theory Connection Two communities use a practical term 'ought' to refer to the P property if: (i) there is a role r and each community uses 'ought' with role r, (ii) r approximates a true theory T, (iii) T refers to P in its law-like generalizations, and **(iv) there are no additional roles r^* and theories T^* such that 'ought' is used with r^* by one of the communities and r^* approximates T^* but not T.**

This is not a complete theory, since it is only a sufficient condition for when two communities manage to refer to the same property. But it does make a general claim which entails that many communities, out of all the possible communities throughout

modal space, do in fact refer to the same property with their term 'ought', and so are capable of genuine disagreements. These include the communities that use their 'ought' with a moral role, or a normative role, but no roles that approximate alternative theories.

Since practical terms are not universally stable, we also need a generalization that concerns when two communities use a practical term, but fail to refer to the same thing. The motivation for discarding **Role-Theory Connection** in favor of **Unique Role-Theory Connection** provides a clue: this can happen when one community uses their term with another role that approximates an alternative theory. (For example: when the Ability Twins use their 'ought' with a role that approximates the theory that characterizes the best state property, they are not referring to obligation.) This can be generalized to the idea that when a community uses a practical term with an additional role that characterizes a distinct subject-matter, there will be pressure to interpret them as speaking about something different. This is captured by **Divergent Role-Theory Connection**

Divergent Role-Theory Connection Some possible communities, c and c^*, use a practical term 'ought' to refer to distinct properties, P and $P*^*$, because: (i) there is a role r and each community uses 'ought' with role r, (ii) r approximates a true theory T, (iii) T refers to P in its law-like generalizations, but (iv) there is an additional role r^* such that c^* uses 'ought' with r^* while c does not, r^* approximates the true theory T^*, and T^* refers to P^* in its law-like generalizations.

Divergent Role-Theory Connection is not a universal claim about every possible community. Rather it makes an existential claim about some possible communities. It says that stability for practical terms does not extend across every possible community

that uses the relevant language, because some of them use their terms with divergent additional roles, and thereby refer to distinct properties. But this is not necessary: use of a practical term with an additional role might, but does not inevitably, constitute a failure of stability for practical terms.

If stability is explained by fit with use of 'ought', plus the eliteness of candidate referents, it is easy to see why practical terms used with additional roles might, but do not inevitably, result in stability failures. Ex hypothesi a term used with additional roles fits multiple elite properties to some extent. For example, a term used with a moral role will fit moral rightness reasonably well. But if it is used with another role that fits the best state property, then there is another elite property, distinct from moral rightness, which is also fit relatively well. Both properties feature in the the generalizations of good theories of different subject-matters; so they are elite, and can be known to be so. Which property 'ought' refers to depends on the details of the relevant degree of fit and, possibly, the degree of eliteness of the candidate referents. We cannot assume that every pair of possible communities that satisfy the conditions of **Divergent Role-Theory Connection** will thereby refer to different properties. But some do.

The reference-magnetic explanation for why **Divergent Role-Theory Connection** is true gives rise to a final possibility that is worth mentioning. In some cases, it will be vague, or indeterminate, whether two possible communities are referring to the same thing. This is because it can be vague which property maximizes fit with use plus eliteness. In fact, there are guaranteed to be such cases. There are pairs of communities that determinately satisfy **Divergent Role-Theory Connection**: for example, take a community c that refers to moral rightness, and a community c^* that refers to the best

state property. Then, there is a possible community c^{**} that is identical to c^{*} in that their use fits moral rightness to some extent, but fits the best state property slightly less well than the use of c^{*}. And there are further possible communities c^{***} whose use fits the best state property slightly less well than c^{**}, and so on. At some point, there will be communities for which it is neither true that the best state property maximizes fit plus eliteness for their use of 'ought', and neither is it true that moral rightness maximizes fit plus eliteness. Reference in these cases will be indeterminate.[173]

I have argued in this chapter that reference magnetism not only can provide an in-principle explanation of the failure of **Universal Disagreement**; it can also give a principled explanation of why we should expect the in-principle explanation to work. Even if eliteness is metaphysically primitive, an account of the epistemology of eliteness can constrain where we should expect, and where we should not expect, to find elite properties. In general, a plausible epistemology of eliteness will rely on the results of good first-order theorizing to deliver predictions about which properties are elite (these are the properties that appear in the generalizations of true first-order theories). Since there are multiple legitimate areas for theorizing in ethics, broadly construed, we should expect that we can come to know about multiple elite ethical properties, broadly construed. We may not actually know which properties these are, if we are not in a position to, at present, know our the relevant theoretical generalizations of ethics. But we appear to know that there are multiple interesting subject-matters for ethics, which can be characterized by alternative theories. So, we are at minimum in a position to know that there are some elite properties that play the role the reference-

[173]On Epistemicist views of vagueness such as that of Williamson (1994), vagueness will extend even further. Since vague facts are *unknowable* on this view, even when one candidate referent in fact does better on the fit-plus-eliteness metric, it may be too close for us to be able to tell that it does, and hence another instance of vagueness in reference, for the Epistemicist.

magnetic theory needs, in order to explain why **Robust Disagreement** is true, and **Universal Disagreement** is not.

Chapter 6.

Disagreement and convergence

The theory outlined in the first five chapters of this book is motivated by the need to explain one kind of fact about disagreement, namely facts about how it is possible for possible communities to disagree with each other at all. The contrast case is one where the communities *talk past one another*, by using words that are similar in virtue of being used with a practical role but nonetheless have different meanings. In typical cases where a community will say 'one ought to ϕ' while a second will say 'one ought not to ϕ', where the 'ought' is used with the same practical practical role in both cases, the explanatory task is to say why these communities genuinely disagree with each other, by referring to the same property with their term 'ought'. But not all: we need to explain why it is possible for some communities to use their 'ought' with a practical role and yet fail to genuinely disagree.

There is a second distinct issue surrounding disagreement. We can take shared reference between users of practical language as a given—that is, assume that it is a fact that such speakers disagree because they are all talking about the same subject-matter—and ask which theories predict that this disagreement would exist. The contrast here is between disagreement and *agreement*. Instead of referring to the same thing with their term 'ought', namely obligation, and making incompatible claims about it, co-referring speakers could have all accepted the *same* claims. If users of practical language are disposed to agree, in suitable conditions, then they *converge*. But convergence appears to be largely absent among users of practical language. In this chapter I will address the apparent absence of convergence, understood roughly as the persistence of disagreement and absence of agreement. Many opponents of realism

have claimed that, in one way or another, the phenomenon presents problems for the realist.

I will not provide a fully general discussion of this issue. Rather my aim is to take the core components of this book that, I have argued, explain the stability of practical terms. The conjunction of these theses provides a natural and realist-friendly elaboration of how we manage to refer to moral and normative properties, how disagreements with these terms are common across different communities in modal space. The theses that explain this include a metaphysical claim: that the moral and normative properties that we use practical terms to refer to are metaphysically elite. The relevant type of metaphysical privilege is common to all of the subject-matters that are a part of reality. Another component is a meta-semantic thesis: that the elite properties, including rightness and obligation, serve as reference magnets. Finally there is an epistemological component: that we can know what these elite properties are on the basis of knowledge produced by ethical theorizing.

There are multiple, related claims about convergence that have been advanced against realist views. I will consider several that are connected to the main themes of this book. One is that convergence is required for co-reference, which I will discuss in §2. A second is that the absence of convergence is incompatible with moral and normative knowledge; I address this claim in §3. And a third is that any realist view is committed to the prediction that convergence obtains; I close by discussing this claim in §4.

I will take these claims in turn, after some brief methodological remarks (§1). My verdict will be that there is at least one type of realist—one who adopts the metaphysical, meta-semantic, and epistemological theses I have outlined here—who should reject

each claim, for principled reasons.

6.1 Methodology

6.1.1 Explanatory burdens

The objections to realism that I discuss in each section of this chapter rest on general claims about convergence. Each claim constitutes an *objection* to realism because (i) it does not appear that convergence holds, but (ii) it appears, according to the objections, that the realist needs to deny this. This chapter grants, for the most part, the objector's claims about the failure of convergence. Moreover, for the most part, I will remain neutral on the precise formulation of the nature of convergence. In some cases it will be necessary to state some relatively precise ideas about what the conditions under which agreement is supposed to be present are supposed to be. But in order to highlight some general themes I will work with a rough-and-ready characterization of the notion.

The responses to convergence-based objections here do not proceed from neutral principles. They rely on the metaphysics, meta-semantics, and epistemology of the view I develop in earlier chapters of this book, and so the responses are not available to realists who reject one or more components of the view. Nonetheless each component has its own intrinsic plausibility. General claims, which hold that *every* realist view ought to accept the premises of a convergence-based argument, find a counterexample in the conjunction of these theses. If an objection relies on an incompatible premise about the metaphysics, meta-semantics, or epistemology of normative properties, it will fail to be fully general by virtue of not targeting the present view.

The various convergence-related problems put opposing pressures on a realist theory. One problem is an explanatory one: can realism explain the failure of conver-

gence? If not, there is a related inference to the best explanation argument against realism. If the view cannot explain why we should expect to see something like the failure of convergence that we actually observe, then we should reject the view in favor of others that offer a better explanation.[174] There are many details that a good inference to the best explanation argument needs to attend to. We don't reject atomic theory just because it fails to explain the pattern of the tides; there are other explanations for the pattern of the tides. Likewise a metaphysics of normative facts that locates them in reality is threatened by an inference to the best explanation argument only if there are no additional factors outside of the metaphysical nature of the normative facts that fail to explain the absence of a convergence phenomenon.

So some convergence worries pressure the realist to accept convergence, and to offer an explanation of it. But there are pressures against the realist view in the opposite direction, suggesting that the view *must* accept convergence, rather than explain its failure. These will be my focus here. Since a good realist theory should explain how we can know the moral and normative facts, and how different communities can co-refer with their practical terms, then, if co-reference is necessary for knowledge and co-reference, the realist will need to show that convergence *does* obtain. My target is the conditionals connecting convergence to reference and knowledge.

6.1.2 What is Convergence?

Two communities *converge* in their use of practical terms just in case they would, under suitable conditions, accept the same practical claims. Convergence is a modal condition: two speakers might *actually* disagree with one another, but nonetheless

[174] Enoch (2008) distinguishes, and rejects, a number of distinct inference to the best explanation arguments along these lines.

converge, because their disagreement would go away if both were to be in the relevant suitable conditions. Thus the claim that A and B converge does not imply agreement between A and B; the claim implies only that A and B agree in certain conditions which need not be actual.

Convergence arguments are arguments which rely on the premise that users of practical language do not, in general, converge. The proponents of convergence arguments aim to show that realist theories fail because they do not predict the appropriate failures of convergence. It is then incumbent on the proponent of a convergence argument to say precisely, in what conditions, convergence would be expected on the realist view, and why convergence, so understood, would be necessary for the realist. Convergence does not require actual agreement, but rather agreement under "suitable conditions"—but what these amount to is a difficult question.[175]

There are, broadly, a few components of the notion of suitable conditions that feature in a convergence condition. A *full information* condition is one: a disagreement does not show the failure of convergence if the disagreement results from one or more parties being factually uniformed. Take two speakers that actually have incompatible normative beliefs—say one believes that giving money directly to the homeless is obligatory, while the other denies this. They might nonetheless come to share the same beliefs if they were to acquire more information, including information about whether giving money directly to the homeless rather than to established homeless charities is more effective. Thus a full information condition is plausibly one part of the counterfactual suitable conditions involved in any convergence claim.

A further condition on convergence involves what we can call "ideal reflective pow-

[175]See for example Hawthorne and Srinivasan (2013) on related issues in the notion of "peer disagreement". Difficulties for related notion of a "shortcoming-free disagreement" in Wright (1992) are also relevant here.

ers".[176] Even if two speakers have the same information, we should not expect agreement among them if one is subject to systematic biases or makes logical mistakes when reaching conclusions. The modal property of convergence, then, holds only between two speakers when they meet the modal condition of agreeing in the nearest worlds where, in addition to having full information, they have ideal reflective powers.

Perhaps experiences provide more than just propositional content that can be represented as information an agent has learned or their ability to reason with that information. If so, the ideal conditions under which agreement is required for convergence might involve identical experience and feedback. For example: take a community that knows that failure to wash hands causes disease, but fails to care about disease and thereby refrains from washing their hands. Their normative judgments are consistent with this behavior: this community does not apply their 'ought' to hand-washing. Nonetheless this community will not persist in their disagreement with communities that do say 'one ought to wash one's hands', since after they have the experience of not washing their hands and getting sick, they will revise their practical judgments in light of their experience of the causal connection between not washing hands and disease. After repeated experience and feedback, this community will come to agree with others that one ought to wash one's hands.[177] Plausibly then we will to add to the components of the ideal conditions in a convergence condition, and include shared and repeated experiences. Two communities will fail to converge only if, after acquiring full information, having ideal reflective powers, *and* having shared experiential feedback, they continue to disagree.

Of course adding conditions to convergence to require too much similarity between

[176] Cf. Schroeter and Schroeter (2013, 4)
[177] Railton (1986, 174 ff) emphasizes this point.

the communities that need to agree, in order for convergence to obtain, risks trivializing the claim that our normative beliefs converge (every community will agree with a community that is exactly identical to it). But this is not, as I have mentioned, a problem I will try to solve here.

These are just a few broad qualifications to a convergence condition. Why should we hold that normative beliefs do in fact converge, in this sense? Some arguments hold that, if there is no convergence, then normative terms do not co-refer. Others hold that the failure of convergence entails that there is no normative knowledge. And some have argued that if convergence fails, then realism itself must be false. I will address each of these claims, focusing on whether each claim is true given the conjunction of theses I have argued for in previous chapters. Since these theses are each independently motivated, it will be a good test for the ideas that convergence is a plausible independent requirement for co-reference or knowledge. I will argue that each claim about convergence fails this test.

6.2 Convergence and co-reference

At the beginning of this chapter I distinguished between two different problems relating to disagreement for the realist. One to explain how it is possible that a variety of different users across modal space are capable of disagreeing with each other at all. Another is to explain why disagreement persists; i.e., why we do not see (or cannot expect to see) agreement among users of practical language. These are distinct explanatory desiderata, but, given some meta-semantic assumptions, they pose a dilemma for the realist. The meta-semantic assumption in question that is that two communities use 'ought' to refer to the same thing only if they converge. The real-

ist is faced with one of two choices. Either the realist must show that the relevant convergence does hold for every possible pair of communities that have genuine practical disagreements. Or, the realist can concede that convergence is not that extensive, and so there are many communities across modal space who do not have substantive disagreements—practical terms are not very stable. The first option makes a strong claim about the psychological profile of users of practical language, and is not superficially very plausible. The second option returns the realist to facing the objections from Moral Twin Earth-style arguments from Chapters 1 and 2.

While positing widespread convergence may seem implausible, many realist theories entail, or at least suggest, that convergence will obtain among a wide range of speakers who refer to obligation. Boyd (1988) holds that reference is a causal relation, and that communities who are talking about the same thing will, owing to the causal properties of the referent, over time come to use their terms in the same way. So, two communities that refer to obligation will, owing to the causal properties of obligation, come to use their term 'ought' in the same way, assuming their 'ought' is causally regulated by the same property. Jackson and Pettit (1996) hold that moral terms refer to the property that satisfies the platitudes of mature folk theory. In these cases uses over time as additional causal influence occurs, or as folk morality better appreciates the platitudes mature folk theory, will come to resemble each other to greater degrees.[178]

These are examples of particular realist theories which hold that co-referring communities can be expected to converge. Since they also allow that some relatively normal possible communities will not converge with their use of practical language, they are vulnerable to the Moral Twin Earth-style arguments raised by Horgan and Tim-

[178]Peter Railton (1986) likewise develops a theory of moral properties which implies that only individuals who would converge after sufficient experience and feedback are talking about the same thing with their moral terms.

mons. Thus there are instances of realist theories that grant the claim that convergence is required for co-reference, and then are forced to concede that the communities in Moral Twin Earth scenarios do not co-refer because they do not converge. The interesting question, however, is whether this is a necessary feature of realism, or only follows from some specific versions of realism.

There are arguments in the literature for this stronger claim. Here is an argument from Merli (2007) for the conclusion that any meta-semantic theory which predicts that two communities are talking about the same thing will converge in some relevant sense:

> If disagreement persists [between two communities in suitably idealized circumstances], the realist is faced with the unenviable task of explaining how speakers are expressing properties that they *don't* see as "what they're getting at" even at the end of the investigative day. [...] We could insist that these idealized yet recalcitrant speakers are saying erroneous things about right and wrong, but a more charitable reading interprets them as simply talking about something else. If they insist on their views, and we insist on ours, despite all the relevant idealizations, it would be difficult to vindicate the idea that we are all linked to or tracking the same property or kind. (Merli, 2007, 304)

Schroeter and Schroeter (2013) give a similar argument, framed in terms of a "descriptivist" theory of reference-determination that takes the descriptions speakers ideally associate with a term as reference-determining: whatever property best satisfies the relevant descriptions is the property the term refers to. A descriptivist account of co-reference, they claim, carries a commitment to convergence:

> [T]he traditional, broadly descriptivist, approach to meaning and reference determination fixes the reference of a predicate by appealing to the speaker's ideal reflective judgments about what that predicate applies to. On such accounts, speakers co-refer with the term 'morally right' only if they would converge on the same verdicts about which actions count as morally right after ideal reflection. (Schroeter and Schroeter, 2013, 2)

Where Merli appeals to the notion of charity, the theoretical commitments of descriptivism play a similar role for Schroeter and Schroeter. The upshot is the same: interpreting non-converging communities as co-referring will, the argument goes, make predictions that are not consistent with the constraints on a good theory of reference-determination. This is a failure to be maximally charitable (Merli), or a failure to be consistent with a broadly descriptivist approach (Schroeter and Schroeter).[179] Either way, convergence is not just a consequence of some specific theories of reference-determination, but is instead a feature that falls out of broad constraints on any good theory.

But these meta-semantic theses are not necessary commitments of realism. A theory that appeals to charity, or satisfaction of descriptive constraints, holds a position that will have the same problems as a view which holds that certain aspects of use are the only factor in determining reference. There are compelling considerations which suggest that any theory along these lines will be unsatisfactory. For instance it is not plausible that use alone, even in conditions where we are "reflectively ideal", will single out sufficiently determinate referents (*cf.* Chapter 3). This is what motivates **Magnetism** as a theory of reference-determination.

For a realist who accepts **Magnetism** as a solution to problems of underdetermination and indeterminacy in meta-semantics, the charity- and descriptivism-based arguments for a convergence condition on co-reference fail. Not every user of practical language, even in "reflectively ideal" conditions, will know which property in the vicinity of her usage of 'ought' is elite. For example, even if happiness-maximization is not the elite property of obligation, some Utilitarian community will treat 'ought'

[179] Loeb (1998) develops a prototype of this argument, although he is less friendly to the idea that convergence holds.

as if it does refer this property. They will say things like 'killing one person in order to save five is obligatory' in many circumstances. If they are metaphysically inclined, they will think that happiness-maximization is elite, but they will be wrong.

The meta-semantic doctrine of reference magnetism says that these speakers are not referring to happiness-maximization. There is another distinct property that is elite and, given that these speakers are using 'ought' with a normative role, also fits their usage of 'ought' reasonably well. Since reference magnetism says that elite properties are easy to refer to, in this case speakers in the Utilitarian community are referring to obligation, and not happiness-maximization. They co-refer with other users of practical language, including non-Utilitarians, who also use 'ought' to refer to obligation.

The combination of a metaphysical commitment to elite normative properties with a meta-semantic view that includes reference magnetism provides a beginning outline of a picture on which co-reference does not require convergence. Convergence is a modal condition, holding that in order to converge, two communities must agree only in those counterfactual worlds where they meet certain idealized conditions. It is not enough for an eliteness-plus-reference magnetism view to show that *actual* speakers can fail to agree. It needs to be possible for speakers to fail to agree as well, even when they are in idealized counterfactual conditions.

The epistemic profile of elite properties gives us the resources to explain why agreement is not guaranteed to be present even between speakers in ideal circumstances. Since obligation is elite, it must be possible to come to know that obligation is elite. But it is not necessary; some do not know which properties are elite, and would not come to acquire such knowledge even if they were to gain more factual information, become better reasoners, etc. In short, they would not believe the correct moral or normative

theory, even if they were in idealized conditions.

In the framework I sketched in Chapter 5, this means that some possible speakers have true beliefs about which normative properties are elite, and do so without being subject to epistemic risk—that is, they could not easily have had a false belief on the topic. This possibility must exist if we are to justify the claim that there are elite moral and normative properties. Other possible speakers fail to have knowledge because, even if they have true normative beliefs, these beliefs are at risk of being false. This possibility must also exist, if convergence does not obtain. That is, the following claims are true:

Possible Knowledge Some possible speakers know the facts about which normative property is elite, through their use of normative theorizing.

Possible Ignorance There are some possible speakers who have false beliefs about normativity, including false beliefs about which properties are the elite normative properties. Moreover, they continue to have false beliefs in ideal circumstances.

Possible Knowledge and **Possible Ignorance** are compatible, given an anti-risk condition on knowledge. I will defend this claim in more detail in the next section: the existence of possible speakers that make **Possible Ignorance** true does not show that every speaker with true beliefs about the eliteness of certain normative properties are at risk of a false belief. For now, it is sufficient to notice that knowledge only requires the absence of false beliefs in *nearby* worlds, and the possibility of being ignorant does not entail that one could easily have been so.

The speakers that satisfy **Possible Ignorance** are *persistently* ignorant: they not only have false beliefs; their false beliefs about normativity do not go away even in ideal

circumstances. They are, however, talking about the same thing as the speakers that satisfy **Possible Knowledge**, in circumstances of the following kind. First, these are circumstances where the elite property of obligation fits the usage of 'ought' by both speakers reasonably well. Owing to the eliteness of obligation, and the non-eliteness of other properties that are pretty good fits with their use, the ignorant speakers refer to obligation. Moreover, their ignorance is not simply due to their lack of information or bad reasoning: even in ideal conditions, they would continue to accept false normative theories, and as a result have false beliefs about which property is the elite property of obligation. The key point is that while simply giving speakers additional information and making them better reasoners is not sufficient to ensure agreement, this needn't also entail that they fail to meet the conditions to refer to obligation. Both the fit and eliteness components in **Magnetism** are satisfied: obligation is elite, and ex hypothesi these speakers use their 'ought' with a normative role, and so obligation fits their use moderately well.

Of course the persistence of disagreement shows that obligation will fail fit with the use of 'ought' in such a community to a substantial degree. A persistent Utilitarian not only uses her normative language by applying 'ought' to non-obligatory acts; in virtue of her persistence she also has a disposition to continue to use 'ought' in these ways. This is an additional aspect of usage that makes to some extent a property other than obligation a better fit with the persistent Utilitarian's use of 'ought'. But it is not decisive in determining reference. And the ways in which the Utilitarian usage fails to fit obligation do not suggest that there is another elite property that is a better fit. These Utilitarians apply their 'ought' to specific acts that are happiness-maximizing, which is distinct from obligation. But it is not an elite normative property, and so is

not very eligible.

This is simply a sketch of why it is possible, on the realist assumptions outlined here, for non-converging speakers to co-refer. Of course some possible mistakes will be corrected as we approach idealized conditions. But we should not expect every mistake in normative theorizing to be corrected in ideal conditions, if they are specified in a reasonable way. An independently motivated package of metaphysical, meta-semantic, and epistemological theses, moreover, entails that this kind of convergence is not necessary among communities that refer to obligation.

Insofar as motivations for a convergence requirement on co-reference start from non-trivial theoretical commitments in meta-semantics, the realist should feel no pressure to hold that convergence obtains, if there are plausible meta-sematnic commitments that do not have the same consequence. Reference magnetism was introduced in Chapters 3 and 4 for reasons that have nothing to do with convergence. In fact, many of the motivations for reference magnetism suggest that, for independent reasons, reference magnetism will do *better* than a charity-based or descriptivist meta-semantics. The fact that a convergence requirement on co-reference fails on a theory which includes reference magnetism is good evidence that claims to the contrary are mistaken.

6.3 Convergence and knowledge

Lack of agreement—that is, the failure of convergence—about practical matters presents another, deeper problem for the version of realism developed here. In responding to arguments that co-reference with 'ought' requires convergence, I appealed to the assumption that we possibly, but do not necessarily, have knowledge of moral

or normative facts. On the assumptions outlined in Chapter 5, this knowledge can produce, but does not entail, knowledge of which moral or normative properties are elite. This is a natural assumption to make, and it is useful to the realist who wishes to hold that convergence is not necessary for co-reference. But this view threatens to be unstable if the absence of convergence is incompatible with knowledge in the first place.

Bennigson (1996) provides a simple argument for the conclusion that, if our normative beliefs do not converge, then these beliefs cannot be knowledge. A more recent version of the argument is found in Tersman and Risberg (2019), which argues as follows:

> If a and b are also in equally good epistemic positions in relation to P, then the fact that they fail to agree illustrates that neither of them has an epistemic position that is good enough to ensure that they could not easily have failed to be correct. For even if one of them in fact has a correct belief about P, there is also one who has failed to acquire such a belief. In other words, if a believes that P and P is true, then the fact that b does not accept P shows that a's epistemic position nevertheless fails to ensure the adherence of her belief. In this sense, it is just a coincidence that she has a true belief about the matter, rather than being less fortunate as b was. [...] The fact that they cannot reach agreement shows that their epistemic capacities are simply not good enough to allow them to determine, in a robust way, whether P is true. Accordingly, on the adherence requirement, neither of them has knowledge about P.[180]

The argument in this quote shares the epistemological assumptions I am making here: that knowledgeable beliefs must be free from the risk of error, in an appropriate sense. The interesting claim is that, if convergence fails, it then follows that the relevant kind of risk is present, even for those who manage to have true beliefs. I will make two points about this argument within the simple framework for knowledge I am assuming here.[181]

[180] Tersman and Risberg (2019, 201-2)

[181] Tersman and Risberg develop more sophisticated epistemological principles, under the heading of their

The first point is that a risk of error, in order to be incompatible with knowledge, has to be an error that could easily have occurred. If convergence fails, then, by definition, it is possible for someone to be in ideal conditions, and yet fail to have the same normative beliefs as others in the same conditions. Given that the normative facts are the same in each world, at least one party has a false belief.[182] So, error is possible. However, the failure of convergence by itself does not guarantee that the error could *easily* have happened. Something is missing from this argument from the failure of convergence to the absence of knowledge. This is evident from the passage quoted above: the first sentence raises the possibility of a false belief *that could easily have happened*, while this qualification is dropped from the subsequent discussion.

But Tersman and Risburg rely on no additional assumptions in their argument. They do characterize the possibility of error as illustrative of the deficiency of our "epistemic capacities", showing that they are not "sufficiently reliable". But the possibility of error does not on its own introduce epistemically threatening risk, and it does not, on its own, show that our faculties are not reliable. Any epistemic capacity could have gone wrong.

A more realistic picture is that, while it is always true, even for those with true normative beliefs, that it is possible that they believe falsely in idealized conditions, in many cases there is no genuine epistemically relevant risk that they do this. For example: ordinary beliefs formed by perception are true but there is always a possibility that the same belief could have been false, as it could have been formed during an

"adherence requirement". Two points are relevant here. First, analogous claims can be made within the more sophisticated framework. And second, insofar as the additional sophistication adds non-trivial complications to the simple picture I am working with here, objections that rely on the more sophisticated picture are vulnerable to a simple response, which rejects the motivations for the sophistication.

[182]This is not a trivial assumption, since many normative facts are contingent, as they depend on other contingent facts. It is common to assume, however, that some basic normative principles are necessary. We can restrict attention to the failure of convergence over such basic normative principles. Then, the failure to possible parties to agree over such principles entails that one party must have a false belief.

episode of hallucination. Most of us still have perceptual knowledge, however, since we are not at risk of hallucinating.

A second point is that nearby errors are not by themselves incompatible with knowledge. Knowledge-preventing errors need, in addition, to be errors that are arrived at by a roughly similar method or process. Someone with a reliable nervous system and an unreliable thermometer can know that it is hot outside by feeling very warm, and believing that it is hot outside on that basis. Even if she could have easily have used her thermometer, and hence could easily have formed a false belief on the basis of the reading of the thermometer, she still has knowledge when she uses her reliable subjective feeling of warmth. This is because the nearby false belief that is produced by the thermometer is formed by a very different process. But the failure of convergence does not imply that one could have had a false belief *by a similar process*.

In fact, it is very likely that, in many cases, convergence failures for normative beliefs will arise because of uses of substantively different epistemic processes by the non-converging. Take, as an example of an irresolvable normative dispute, a case where a Utilitarian holds that it is permissible to sacrifice the well-being of one in order to improve the well-being of a number of others. A non-convergent speaker who follows Ross (1930), by contrast, settles the question by weighing the prima facie duties that bear on the case. While each party will continue to disagree with the other even in ideal conditions, they do not use the same process to form their beliefs: the Utilitarian engages in a calculation of the net utility produced by the available actions, while the other applies the principles that govern prima facie duties. These are potentially very different kinds of reasoning, and hence different belief-forming processes. This has consequences for the relationship between convergence failures and the absence of

knowledge.

There is a strong analogy with cases where someone has two very different methods available to them, only one of which is reliable. A deontologist could, in principle decide to use calculations of the overall utility produced by each option, in order to decide which act is required. Or, the deontologist could apply principles about prima facie duties. Even if each method is available, they are very different procedures for determining which action is required. There are at least three reasons for this: (i) the content of the principles are distinct; (ii) the principles are structured in different ways, as the Utilitarian has one general principle to apply, whereas the Rossian deontologist has many, with no master principle to determine how the principles are to be weighed against each other in cases of conflict, and (iii) acceptance of the relevant principles engage very different conative mechanisms: someone who accepts a prima facie duty not to break a promise might do so on the basis of feelings of guilt at the prospect of being disloyal. But application of a utility-maximizing principle makes will not engage such a system.[183]

These points are relevant to a related worry that appears in Rowland (2017).[184] If convergence fails, then it is possible to be in the position of someone who faces something like a *peer disagreement.* In typical cases of peer disagreement, there are believers who are in the same epistemic position, but who arrive at different judgments. (One paradigmatic example of peer disagreement is from Elga (2007): two people, watching the finish of a horse race from identical positions, might disagree about which horse won.) If one's peer's judgments are inconsistent with one's own, then, Rowland's ar-

[183]Utilitarians will treat the prospect of producing guilt as part of the consequences of an act that detract from the amount of overall utility that it produces. But this does not mean that the guilt-tinged feelings constitute any part of the acceptance of any normative principles, for the Utilitarian.
[184]Vavova (2014) gives a separate response to arguments of this kind.

gument goes, we should "concilitate"–i.e., reduce confidence in our own judgment. Those who should reduce their own confidence, plausibly, cannot *know* the claim they have reduced confidence in.[185] So, conciliation is incompatible with retaining knowledge. If the failure of convergence of normative belief requires us to conciliate, we cannot have normative knowledge.

This is just a sketch of the kind of worry Rowland raises for the relationship between convergence failures and knowledge. But with this sketch in hand, we can see how the points made above are relevant. First, peer disagreement is more worrisome, epistemically speaking, when we either encounter actual peers who disagree, or are aware of possible peers who could easily have disagreed with us. Knowing that there is a bare metaphysical possibility of a disagreeing peer does not generate much pressure to change one's views. Plausibly, it generates no pressure at all. The difference is one of distance across modal space, and an anti-risk condition on knowledge treats nearby false beliefs very differently from merely possible false beliefs. Only the former subject a belief to epistemic risk. As I argued above, the failure of convergence does not on its own entail that there are nearby agents who share our epistemic position and disagree with us. Similarly it does not entail that disagreeing peers could easily have existed. Rather all it entails is that there is a bare metaphysical possibility that someone in a similar epistemic position disagrees.

6.4 Predictions for convergence?

There may still be a lingering sense that realism requires, for some reason or another, a convergence condition. I have argued that it does not require convergence in order to predict an appropriate range of co-reference between users of practical lan-

[185] Though see Lasonen-Aarnio (2010).

guage. Moreover a realist view does not need convergence in order to accommodate the epistemic claim that we can know some moral or normative claims. But some have argued for a close relationship between convergence and realism without relying on ancillary claims about co-reference or knowledge.

There are various ways to flesh this idea out. Some begin with the idea that other objective domains appear to satisfy a convergence condition: there is no disagreement over whether my desk is solid, whether water boils at 100 degrees Celsius, etc. More carefully, there is near universal agreement about these matters in ideal conditions and, in the cases where we do not agree over these objective facts even in ideal conditions, there is an explanation of why we do not agree. But in normative matters there is nothing like agreement to a significant extent, and (the argument goes) the extent of disagreement cannot be explained if realism is true.

Mackie (1977, 36) presents an "argument from relativity" that is a proto version of an explanatory argument of this kind, and the theme has been taken up by Loeb (1998) and Leiter (2002). If the issue is one of explaining the extent of disagreement, then there are a number of confounding factors for explanatory arguments against realism of this kind.[186]

But there is a more basic point, which I will make in closing. This chapter began with a contrast between two problems about disagreement: one of explaining how disagreement is possible, and another of explaining why an amount of agreement exists alongside disagreement. The first five chapters of this book were occupied with an explanation of the first problem; this chapter deals with the second. But they are not unrelated, and not just because both are problems about "disagreement" at some

[186]Cf. Enoch (2009, §4)

level. As Enoch (2009, 27-8) notes, an explanation of the extent of disagreement is not the only potential virtue of a realist metaphysics of the normative. We might have other reasons to accept such a theory. If so, then the mere fact that the distinctive parts of the realist's metaphysics do not enter into an explanation of one specific data point is not necessarily a mark against the view. Realism is not a theory of everything.

The realist can strengthen this point. Let us grant that there is no specific explanation, from the realist's metaphysics, to the details of the ways in which normative beliefs fail to converge. Still, it is a presupposition of this datum that in cases where convergence fails, we do in fact *disagree*, rather than talk past one another. The realist has a very good explanation for this fact, because reference magnetism is a very natural meta-semantic thesis for the realist to adopt. This is arguably a more central explanatory task than the project of deriving any specific mechanism that explains the ways in which we fail to agree from realist principles.

We can illustrate this with a very simple model of the resources the realist's opponent can use to explain why we fail to converge on moral matters. Suppose convergence fails because, even in ideal conditions, some people fail to come to the same moral conclusions as others since they have a personal interest in seeing themselves as good people.[187] This explanation suggests that there is a possible scenario where someone in ideal conditions applies their moral term 'right' to a happiness-maximizing act, because they did something that had good consequences, and are retrospectively inclined to formulate moral beliefs that entail that they are a good person. It also follows that there is a possible scenario where someone in ideal conditions applies their moral term 'right' to a non-happiness-maximizing act, for analogous reasons.

[187]Loeb (1998, 283). Since people in different circumstances will inevitably do different things, having an interest in seeing oneself as a good person will inevitably lead to opinions on moral matters that differ from the opinions of those in different situations.

What this simple model does not explain is that the failure of these agents to use 'right' in the same way, even in ideal circumstances, is a disagreement about moral matters. Since they use 'right' differently, it follows that they form different moral beliefs. But it does not follow that in both cases the moral beliefs about rightness—i.e., that their differences amount to a genuine disagreement. The point can be repeated for more sophisticated accounts of why convergence fails, which appeal to resources an anti-realist would accept. Even if the failure of all speakers in ideal conditions to use their moral term 'right' in the same way can be explained using these resources, the fact that it is a genuine disagreement does not follow from this explanation alone.

This is where the realist's metaphysics is a helpful, and perhaps necessary, explanatory resource. The first five chapters of this book sketch the ways in which the realist's metaphysics of elite properties, coupled with plausible meta-semantic principles, explains why in cases where speakers use their practical terms differently, they are still talking about the same thing. We should think of this as the distinctive realist contribution to an explanation of failures of convergence. Even if the view has nothing distinctive to say about why speakers sometimes fail to converge, it is essential for explaining why, in these cases, the failure of convergence is an instance of a genuine disagreement. Very plausibly, this is a fact that any meta-ethical theory should explain. The realist can claim that she has the best explanation, and does not need to punt to entirely unrelated issues in meta-ethics to claim a superior explanatory profile for a realist metaphysics of moral and normative properties.

Non-cognitivist accounts of disagreement, by contrast, risk overgeneralizing. If the account of why there is genuine disagreement between non-convergent speakers appeals only to the fact that they make superficially contradictory claims using terms

that share a practical role, then the non-cognitivist position predicts **Universal Disagreement**. I have not developed this argument in detail here. But the methodological point is clear: realism can provide an explanation of a whole range of facts involving disagreement. This involves (i) showing that plausible principles in metaphysics, meta-semantics, and epistemology entail that disagreement with practical terms is robust, but not universal, in the technical senses of "robust" and "universal" that I have used here, and (ii) deferring to ancillary facts about psychological constitution to explain why not every possible speaker arrives at true beliefs about the referents of practical terms in their own language. A final evaluation of this view would require a comparative claim about the explanatory power of its competitors. I have not provided the materials for a full evaluation here. But it seems unlikely that sniping at the realist view using specific claims about convergence will succeed. Instead the task is to develop a competing non-realist view that does as well or better than the realism sketched here at explaining a whole range of central explanatory desiderata for a theory in meta-ethics, including facts about disagreement.[188]

[188] As urged in Williamson (2006).

Conclusion.

Eklund's Bad Guy raises questions that are related to the semantic stability of practical terms, but highlights some important consequences of stability that go beyond mere facts about disagreement. If Bad Guy should be interpreted as using his normative 'ought' to refer to obligation—the same property we refer to—then it follows that there is a measure of semantic stability for 'ought'. If not, then Bad Guy can say things like 'one ought not give money to the poor', using an 'ought' with a normative role, and say something that is *true* in his language. In closing I will summarize the main claims of this book and then briefly discuss what the realist view that accepts them has to say about Bad Guy.

Chapter 1 sketched a natural generalization of traditional examples of disagreement with practical terms. Not only do the speakers on Horgan and Timmons's Earth and Moral Twin Earth appear to disagree with one another, the example can easily be extended to other Moral Twin Earth-style communities who appear to disagree in the same way. This suggests the hypothesis: *every* possible community that uses a practical term with the same moral or normative role is capable, no matter what other differences there may be between them, of having a substantive disagreement.

This generalization, while tempting, is not the only one available. There are different kinds of differences. In the original Moral Twin Earth case and variants from Chapter 1, the disagreeing communities differ in what kinds of actions they hold to be right, or obligatory. It is natural to describe them as differing only in which substantive theory of the relevant subject-matter they accept. There are, however, other possible differences. What makes it natural to describe certain communities as those that accept a particular theory of moral rightness, or all-things-considered obligation, is the

role with which they use their terms 'right', 'ought', and the like. But some possible speakers use their moral or practical terms with additional roles which characterize other morally or normatively significant properties. In Chapter 2, I argued that in some instances these speakers will intuitively not be disagreeing with those who use their practical terms to refer to moral rightness and all-things-considered obligation.

The upshot is that shared moral or normative role is not sufficient for the capacity to disagree. A possible community can use their 'ought' the the same normative role that we use ours with, treating it as immediately guiding action in deliberation. But, if they are using this 'ought' with additional roles that clearly characterize other normatively interesting properties, we are inclined to say that they are referring to something other than obligation. They do not disagree with us. Cases like this set a clear explanatory target: we need to explain why communities like these do not disagree with communities who use their practical terms to refer to the usual properties. But at the same time we need to do this without conceding that every significant difference in usage of a practical term results in the same kind of divergent reference. We still need to explain what the original Moral Twin Earth cases illustrate, which is that practical terms are stable across a wide range of possible uses. One characterization of the explanatory challenge is that we need a theory that explains **Robust Disagreement** without entailing **Universal Disagreement**.

Chapters 3 and 4 provide such an explanation using resources that are especially friendly to a realist about morality and normativity. The first resources is a metaphysical one, namely the idea that some properties are metaphysically elite: these are the the properties that confer genuine similarity, feature in laws, and are projectable. In general, the fact that these properties are elite implies that they play an important ex-

planatory role in relation to a number of objectively significant features of the world. Realism, as I have characterized it, holds that moral and normative properties are among the elite.

The second resource I appealed to is reference magnetism, which is an account of what makes it the case that terms in a language refer to particular parts of the world. Since reference magnetism holds that some terms refer to certain objects or properties, rather than others, in part because of the eliteness of their referents, it relies on a distinctive feature of realism about morality and normativity. Moreover I argued that it is defensible both as a general theory of reference-determination, and as a theory of reference-determination for practical terms. Most importantly, it can explain why **Robust Disagreement** is true, without generalizing into a commitment to **Universal Disagreement**.

This explanation relies on certain assumptions about which moral and normative properties are elite. Importantly, it assumes that there are multiple elite moral properties (and similarly for normative properties), but not too many. (If there are many elite moral properties, then moral terms will not be sufficiently stable to support **Robust Disagreement**.) The primitivism about eliteness of Chapters 3 and 4 precludes a kind of direct explanation of why the elite properties would be distributed in the required way. Chapter 5 provides an indirect explanation. Since, according to one component of the eliteness-role, elite properties show up in law-like generalizations, we can exploit a plausible account of how we *know* about the relevant generalizations. The construction of a theory, plausibly, is a way to discover what the law-like generalizations about a particular subject-matter are. Theory-construction in ethics, then, is plausibly a way to discover—that is, come to know—which ethical properties are elite. This can be

revealed by which properties show up in the law-like generalizations of these theories.

The contribution of Chapter 5 is the conjecture that there are multiple elite moral properties, on the grounds that there are multiple distinct areas of theoretical investigation that are of interest to ethics. One interesting area for investigation is the familiar question of what we morally ought to do, in the familiar sense. In principle it is possible to know the theory that answers this questions, and this theory (on present assumptions) includes generalizations that refer to the elite property of moral obligation. But this is not the only interesting theory: we might also be interested in what the best way for the world to be is, regardless of the contingent abilities of anyone to bring it about. Other areas of theoretical investigation concern what the principles that limited agents like us should use in decision-making, and the morally required actions that agents can realistically be expected to perform. If these are legitimate areas for theorizing, then the properties in Chapter 2—the best state property, the property of being boundedly obligatory, and the property of psychologically feasible obligation—can be known to be elite.

This has an immediate consequence for the counterexamples to **Universal Disagreement**, given reference magnetism as a theory of reference-determination: some possible communities that use their practical terms to fit with these elite properties, which are distinct from obligation, will be referring to the best state property, or the property of being boundedly obligatory, or the property of psychologically feasible obligation. It is not guaranteed that *every* possible use of a moral term is such that the elite property of moral rightness maximizes the fit and eliteness components that are relevant to reference-determination. There are examples of other elite, morally relevant properties, which in some possible cases do better on the fit component. This

also yields a prediction, from within the theory presented here: there will be other analogous counterexamples to **Universal Disagreement** in the form of other possible communities that use their practical terms with roles that characterize other theoretically interesting properties that are of practical relevance. In cases like this, if the theory I have outlined here is right, we should have similar intuitions of an absence of substantive disagreement.

The central explanation I am offering here is indirect in one straightforward way: given primitivism about eliteness, there is nothing that makes these morally relevant properties elite; it is a primitive fact that they are elite. Theoretical generalizations only provide an epistemic resource for *knowing* that these properties are elite. It is indirect in a second way: the explanation makes no commitment to the claim that we in fact know the theory which contains the true generalizations about moral rightness (or any other practically relevant elite property), and moreover is not committed to the claim that we will come to know these generalizations, even under idealized conditions. As I outline in Chapters 5 and 6, given some plausible structural features of knowledge, all of this is consistent. We can provide the schematic explanation of why we should expect there to be failures of **Universal Disagreement**, without actually knowing what the morally and normatively relevant properties are.

To conclude I will sketch what lessons we can take from this picture for Eklund's Bad Guy. The thought experiments that raise to salience generalizations such as **Robust Disagreement** and **Universal Disagreement** are in the first instance about the disagreement-relations between linguistic communities throughout modal space. Given the additional assumption (which I claimed in Chapter 1 the realist should accept) these disagreements are *substantive* disagreements, we should explain whatever

generalization about disagreement across modal space is true in terms of the semantic stability of practical terms. That is, the question of whether **Robust Disagreement** or **Universal Disagreement** is true boils down to the question of whether **Robust Stability** or **Universal Stability** is true.

Eklund's Bad Guy, by contrast, raises the issue of semantic stability by another route. The worry is that, if the realist must concede that "he just does not employ our notion of reason or our notion of what ought to be done but instead employs alternative normative notions" (Eklund, 2017, 5) since he refers to a different property than we do with his normative terms, then the options for saying how Bad Guy is objectively mistaken are severely restricted. By employing an "alternative normative notion", Bad Guy doesn't say anything false, in his language. The worry in this case can be framed as a worry about the objectivity of normativity.[189] There is no straightforward diagnosis, that both we and Bad Guy would accept, of where he goes wrong. If he says 'one ought to steal from the poor', *we* say that he says something false, but in *his* language, 'one ought to steal from the poor' is true. The mistake, if any, is much more subtle.

However if practical terms are semantically stable—and hence Bad Guy means what we mean with his normative 'ought'—the problem goes away. When he says 'one ought to steal from the poor' he says something false, since ex hypothesi Bad Guy's 'ought' refers to obligation, and stealing from the poor does not instantiate obligation. So if reference magnetism can explain why a normative 'ought' is stable between our case and Bad Guy's, the question of why Bad Guy is mistaken is easily answered. I won't rehearse here how the realist can appeal to reference magnetism to deliver this

[189] As Eklund's extended discussion makes clear, he doesn't think that the absence of stability *entails* that normativity isn't objective. But it is equally clear that he thinks there is no obviously satisfactory explanation of the objectivity of normativity, if normative terms are not stable.

answer, or why Eklund's objection to reference magnetism is mistaken.

What is worth emphasizing is that the objectivity worries do not press us to endorse **Universal Stability**, for reasons related to the import for stability for practical terms from thought experiments about disagreement. Consider a version of Bad Guy who, instead of using a normative term which is identical to ours with the exception that he applies it to acts that are wrong, differs from us in the way the Ability Twins differ. That is: this version of Bad Guy—which we can call *Ability Bad Guy*—uses his 'ought' with a normative role but, in addition, associates with his 'ought' an additional role. This role in effect holds that when 'ought' is applied to a subject and an action, it is not necessary that the subject, with whatever contingent abilities and powers they happen to have, be in a position to be able to perform the action in question.

For the same reasons as before, it is very natural to interpret Ability Bad Guy as speaking about what I dubbed the *best state property*, which is distinct from obligation. But in this case, there is no apparent threat to the objectivity of normativity for the realist view I have sketched.[190] While there is something bizarre about Ability Bad Guy's use of a term with a normative role to speak about the best state property, we can understand him as referring to the best state property, and so speaking truly when he says 'Sally ought to relieve a famine', even though Sally has no ability to do so in her present circumstance. This does not mean that we cannot say that there is no objective fact about whether Sally ought to relieve the famine. Instead we say that, while Sally is not currently obligated to end the famine, since she has no ability to do so, it would in fact be best if she did. We are stating the former fact with our use of 'ought', while Ability Bad Guy states the latter. Showing that a generic Bad Guy is wrong is only one

[190] I leave it as a open question what the Ardent Realism, as Eklund defines the view, should say about Ability Bad Guy.

way to maintain the objectivity of normativity. Showing that he is saying something true about another legitimate area of normative theorizing is another.

Ability Bad Guy is very different from Eklund's original Bad Guy in this respect, according to the realist view I have outlined. Ability Bad Guy is speaking about an elite property that is highly important for moral and normative theorizing. It is not, I am assuming, a property that is identical to what we refer to with our term 'ought', since what we refer to bears significant connections to what a subject is able to do. Not only is it important; it is one *we* should objectively recognize as such. This is an important contrast with Bad Guy in his original form. This version of Bad Guy applies his normative 'ought' to actions that only false normative theories claim are obligatory. Acknowledging that he uses his normative terms to say true things using his normative terms would indeed be worrisome for objectivity-related reasons for a realist. It is not obligatory to steal from the poor (it is obligatory not to steal), and so it would be prima facie bizarre if someone who simply used their normative 'ought' differently than we do managed to speak truly when they say 'one ought to steal from the poor'. Making true normative statements does not simply depend on how one chooses to use one's normative terms.

Things do not look the same when we consider possible speakers who apply their practical terms to other elite moral or normative properties. While there is something bizarre about the speaker who applies their normative 'ought' to the best state property,[191] they are referring to something that does in fact matter, practically speaking. If both we and Ability Bad Guy are speaking about different, highly elite normative properties, there is no obvious cause for concern. After all, these are all metaphysically

[191]They will end up judging themselves as unavoidably incoherent, since they apply 'ought' to actions and do not perform these actions—see Chapter 2.

privileged and normatively relevant subjects. It is not in general true that using one's normative terms differently suffices for a change in subject-matter, but using one's normative terms to fit with a property that is both theoretically significant for moral theorizing and highly elite can suffice.

Cases like this do generate a measure of paradox. By definition, the term 'ought' as used by a community is normative just in case that community uses their 'ought' with a normative role. I have argued that a speaker like Ability Bad Guy might use his normative 'ought' to refer to something besides obligation—for example, Ability Bad Guy might refer to the best state property. Thus, he uses his normative term 'ought' to refer to the best state property, which is not the normative property of obligation. It might seem paradoxical to hold that a normative 'ought' in the mouths of some possible speaker doesn't refer to the normative property of obligation. But this is an artifact of our stipulative definition of 'normative term', one on which 'ought' is a normative term just in case it is used with a normative role.

I will not try to articulate what makes a property normative here.[192] But if normative role is thin—for instance, if a normative role is characterized as the Gibbard role—then it is natural to say that the fact that a property is referred to by a normative term is not sufficient to make it a normative property. This is because obligation is a normative property, but it is possible that 'ought' is used with the Gibbard role and does not refer to obligation. Of course there are options here: we could supplement our definition of 'normative role' to include a "thick" role that secures a single referent across all possible uses. Or we could widen the definition of a normative property to include normatively relevant properties, such as the best state property.

[192]This question is raised in Eklund (2017, Chs. 4-5).

This is one question among many that is left open by what I have argued for in this book. I haven't given a complete realist theory of morality and normativity here; rather, I have sketched the core components of one, which is developed around a few core commitments. These are: a metaphysics of elite properties; a meta-semantics that includes reference magnetism, and an anti-risk epistemology. This package of views does not constitute, for all I have argued here, the only viable option for the realist. I have only argued that it is a natural, defensible, and in some respects promising route for the realist.

The central respect in which this version of realism is promising, I have argued, is in its account of disagreement: in particular, it provides a natural and adequate account of which possible communities that use practical terms are capable of disagreeing with each other. Existing literature emphasizes the extent of disagreement with practical terms across moral space; or, more fundamentally for the realist, the semantic stability of practical terms. But equally important for explanatory purposes, but less central to existing discussions, is the limits to disagreement (or stability). The primary benefit of the realist view I have described here is that it can explain both.

Why is it that practical terms are highly stable? And why, in particular, do communities that use their terms with the same role, but apply them to different actions, appear to disagree with each other? This is because their is a single elite property that fits well enough with the use of the relevant practical term in both communities—notwithstanding the fact that this elite property will fail to be a perfect fit for at least one community. And why are practical terms not stable across all uses with a shared practical role? This is because there are some possible communities that use their practical terms with additional roles, which characterize the subject-matter of other areas

of morally or normatively relevant theorizing. The properties involved in such subject-matters fit the use of some of these communities fairly well. They are also elite, and so provide eligible referents that are distinct from the ordinary referents of our practical terms. There are limits to stability.

The realist who accepts the outline of the view I have provided here can explain both the extent and limits of this phenomenon. While disagreement and related semantic phenomena are often cited as areas of difficulty for the realist, proper attention to a full characterization of what needs to be explained, plus utilization of the right resources, can mitigate the challenge. If competing views have difficulty explaining the full range of data, issues surrounding meta-semantics and disagreement might even be turned into an advantage for the realist.

Bibliography

D. M. Armstrong. *What is a Law of Nature?* Cambridge Univsersity Press, 1983.

Ralf Bader. The Grounding Argument against Non-reductive Moral Realism. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics vol. 12*, pages 106–134. Oxford University Press, 2017.

Thomas Bennigson. Irresovable Disagreement and the Case Against Moral Realism. *Southern Journal of Philosophy*, 34(4):411–437, 1996.

Gunnar Björnsson and Tristram McPherson. Moral Attitudes for Non-Cognitivists: Solving the Specification Problem. *Mind*, 123(489):1–38, 2014.

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

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

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

David O. Brink. *Moral Realism and the Foundations of Ethics*. Cambridge Univsersity Press, 1989.

David O. Brink. Realism, Naturalism, and Moral Semantics. *Social Philosophy and Policy*, 18(2):154–176, 2001.

John Broome. Normative Requirements. *Ratio*, 12(4):398–419, 1999.

John Brunero. Instrumental Rationality, Symmetry, and Scope. *Philosophical Studies*, 157:125–140, 2012.

Herman Cappelen and John Hawthorne. *Relativism and Monadic Truth*. Oxford University Press, 2009.

Matthew Chrisman. *The Meaning of 'Ought'*. Oxford University Press, 2015.

John Conlisk. Why Bounded Rationality? *Journal of Economic Literature*, 34(2):669–700, 1996.

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

Stephen Darwall. *The Second Person Standpoint: Morality, Respect, and Accountability*. Harvard University Press, 2006.

Cian Dorr and John Hawthorne. Naturalness. In Karen Bennett and Dean Zimmerman, editors, *Oxford Studies in Metaphysics vol. 8*. Oxford University Press, 2013.

J. L. Dowell. A Flexible Contextualist Account of Epistemic Modals. *Philosophers' Imprint*, 11(14):1–25, 2011.

J. L. Dowell. The Metaethical Insignificance of Moral Twin Earth. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics, vol. 11*. Oxford University Press, 2015.

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

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

Billy Dunaway. Expressivism and Normative Metaphysics. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics, vol. 11*. Oxford University Press, 2016.

Billy Dunaway. Review of *The Meaning of 'Ought': Beyond Descriptivism and Expressivism in Metaethics* by Matthew Chrisman. *The Journal of Philosophy*, 114(3):155–159, 2017a.

Billy Dunaway. Luck: Evolutionary and Epistemic. *Episteme*, 14(4):441–461, 2017b.

Billy Dunaway. Realism and Objectivity. In Tristram McPherson and David Plunkett, editors, *Routledge Handbook of Metaethics*. Routledge, 2017c.

Billy Dunaway. Epistemological Motivations for Anti-Realism. *Philosophical Studies*, 175(11):2763–2789, 2018.

Billy Dunaway. The Metaphysical Conception of Realism. MS.

Billy Dunaway and Tristram McPherson. Reference Magnetism as a Solution to the Moral Twin Earth Problem. *Ergo*, 3(25):639–679, 2016.

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

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

Matti Eklund. *Choosing Normative Concepts*. Oxford University Press, 2017.

Adam Elga. Reflection and Disagreement. *Noûs*, 41(3):478–502, 2007.

David Enoch. How is Moral Disagreement a Problem for Realism. *Journal of Ethics*, 13: 15–50, 2009.

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

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

Stephen Finlay. Oughts and Ends. *Philosophical Studies*, 143(3):315–340, 2009.

Stephen Finlay. Disagreement Lost and Found. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics vol. 12*, pages 187–205. Oxford University Press, 2017.

Philippa Foot. Moral Beliefs. *Proceedings of the Aristotelian Society*, 59:83–104, 1958/9.

Philippa Foot. Morality as a System of Hyopthetical Imperatives. *The Philosophical Review*, 81(3):305–316, 1972.

Jane Friedman. Suspended Judgment. *Philosophical Studies*, 162(2):165–181, 2013.

P. T. Geach. Whatever Happened to Deontic Logic? *Philosophia*, 11:1–12, 1982.

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.

Nelson Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, 1955.

Sven Ove Hansson. Science and Pseudo-Science. *Stanford Encyclopedia of Philosophy*, 2014.

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

John Hawthorne. A Corrective to the Ramsey-Lewis Account of Theoretical Terms. *Analysis*, 54:105–110, 1994.

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.

John Hawthorne and Amia Srinivasan. Disagreement without transparency: Some bleak thoughts. In David Christensen and Jennifer Lackey, editors, *The Epistemology of Disagreement*. Oxford University Press, 2013.

Avram Hiller and Ram Neta. Safety and Epistemic Luck. *Synthese*, 158:303–313, 2007.

Eli Hirsch. *Dividing Reality*. Oxford University Press, 1997.

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

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

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

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

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 and Philip Pettit. Moral Functionalism, Supervenience and Reductionism. *The Philosophical Quarterly*, 46(182):82–86, 1996.

Niko Kolodny. Why Be Rational? *Mind*, 114:509–563, 2005.

Angelika Kratzer. What "Must" and "Can" Must and Can Mean. *Linguistics and Philosophy*, 1:337–355, 1977.

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

Imre Lakatos. Science and Pseudoscience. In G. Vesey, editor, *Philosophy in the Open*. Open University Press, 1974.

Maria Lasonen-Aarnio. Unreasonable Knowledge. *Philosophical Perspectives*, 24:1–21, 2010.

Stehpanie Leary. Non-naturalism and Normative Necessities. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics vol. 12*, pages 76–105. Oxford University Press, 2017.

Brian Leiter. *Nietzsche on Morality*. Routledge, 2002.

David Lewis. *Convention: A Philosophical Study*. Harvard University Press, 1969.

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

David Lewis. Languages and Language. In Keith Gunderson, editor, *Minnesota Studies in the Philosophy of Science 7*, pages 3–35. 1975.

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.

David Lewis. Meaning without Use: A Reply to Hawthorne. *Australasian Journal of Philosophy*, 70(1):106–110, 1992.

David Lewis. Humean Supervenience Debugged. *Mind*, 103(412):473–490, 1994.

Don Loeb. Moral Realism and the Argument from Disagreement. *Philosophical Studies*, 90(3):281–303, 1998.

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.

David Manley. Keeping Up Appearances: A Reducer's Guide. *The Journal of Philosophy*, forthcoming.

David Manley. Moral Vagueness, Twin Earth, and Semantic Plasticity. MS.

Tristram McPherson. What is at Stake in Debates Among Normative Realists? *Noûs*, 49(1):123–146, 2015.

Tristram McPherson. Authoritatively Normative Concepts. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics vol. 15*. Oxford University Press, 2015.

Tristram McPherson. Ardent Realism without Referential Normativity. *Inquiry*, 2018.

David Merli. Moral Convergence and the Univocity Problem. *American Philosophical Quarterly*, 44(4):297–313, 2007.

David Merli. Expressivism and the Limits of Moral Disagreement. *Journal of Ethics*, 12: 25–55, 2008.

G.E. Moore. *Principia Ethica*. Cambridge University Press, 1903.

Lauren Olin. Questions for a Theory of Humor. *Philosophy Compass*, 11(6):338–350, 2016.

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

Duncan Pritchard. *Epistemic Luck*. Oxford University Press, 2004.

Hilary Putnam. Meaning and Reference. *Journal of Philosophy*, 70(19):699–711, 1973.

Hilary Putnam. The Meaning of 'Meaning'. In *Mind, Language, and Reality: Philosophical Papers, vol. 2*, pages 215–271. Cambridge University Press, 1975.

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

W. V. O. Quine. *Word & Object*. The Mit Press, 1960.

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

Debbie Roberts. Shapelessness and the Thick. *Ethics*, 121(3):489–520, 2011.

Jacob Ross. The Irreducibility of Personal Obligation. *Journal of Philosophical Logic*, 39: 307–323, 2010.

W. D. Ross. *The Right and the Good*. Clarendon Press, Oxford, 1930.

Richard Rowland. The Significance of Fundamental Moral Disagreement. *Noûs*, 51(4): 802–831, 2017.

Jonathan Schaffer. Two Conceptions of Sparse Properties. *Pacific Philosophical Quarterly*, 85:92–102, 2004.

Mark Schroeder. Means-end Coherence, Stringency, and Subjective Reasons. *Philosophical Studies*, 143:337–364, 2009.

Mark Schroeder. *Noncognitivism in Ethics*. Routledge, 2010.

Mark Schroeder. Ought, Agents, and Actions. *The Philosophical Review*, 120(1):1–41, 2011.

Laura Schroeter and François Schroeter. Normative Realism: Co-Reference without Convergence? *Philosophers' Imprint*, 13(13):1–24, 2013.

Wolfgang Schwarz. Against Magnetism. *Australasian Journal of Philosophy*, 92(1):17–36, 2014.

Kieran Setiya. *Knowing Right from Wrong*. Oxford University Press, 2012.

Russ Shafer-Landau. *Moral Realism: A Defense*. Oxford University Press, 2003.

Theodore Sider. *Writing the Book of the World*. Oxford University Press, 2012.

Alex Silk. *Discourse Contextualism: A Framework for Contextualist Semantics and Pragmatics*. Oxford University Press, 2016.

Alex Silk. Normative Language in Context. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics vol. 12*, pages 206–243. Oxford University Press, 2017.

Herbert A. Simon. Theories of Bounded Rationality. In C. B. McGuire and Roy Radner, editors, *Decision and Organization*, pages 161–176. North-Holland Publishing Company, 1972.

Michael Slote and Philip Pettit. Satisficing Consequentialism. *Proceedings of the Aristotelian Society*, 58:139–176, 1984.

Michael Smith. *The Moral Problem*. Wiley-Blackwell, 1994.

Scott Soames. *Beyond Rigidity: the Unfinished Semantic Agenda of Naming and Necessity*. Oxford University Press, 2002.

Ernest Sosa. How to Defeat Opposition to Moore. *Philosophical Perspectives*, 13:141–153, 1999.

C.L. Stevenson. The Emotive Meaning of Ethical Terms. *Mind*, 46:14–31, 1937.

Bart Streumer. Are There Irreducibly Normative Properties? *Australasian Journal of Philosophy*, 86(4):537–561, 2008.

Bart Streumer. *Unbelievable Errors*. Oxford University Press, 2017.

Sarah Stroud. Conceptual Disagreement. *American Philosophical Quarterly*, 56(1):15–28, 2019.

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.

Jussi Suikkanen. Non-naturalism and Reference. *Journal of Ethics and Social Philosophy*, 11(2):1–24, 2017.

Christine Tappolet. The Normativity of Evaluative Concepts. In A. Reboul, editor, *Mind, Values, and Metaphysics*, pages 39–54. Springer, 2014.

Folke Tersman and Olle Risberg. A New Route from Moral Disagreement to Moral Skepticism. *Journal of the American Philosophical Association*, 2(5):189–207, 2019.

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.

Katia Vavova. Moral Disagreement and Moral Skepticism. *Philosophical Perspectives*, 2014.

Brian Weatherson. What are Good Counterexamples? *Philosophical Studies*, 115(1):1–31, 2003.

Brian Weatherson. The Role of Naturalness in Lewis's Theory of Meaning. *Journal of the History of Analytical Philosophy*, 1(10):1–19, 2012.

Ralph Wedgwood. Conceptual Role Semantics for Moral Terms. *The Philosophical Review*, 110(1):1–30, 2001.

Ralph Wedgwood. The Meaning of 'Ought'. In Russ Shafer-Landau, editor, *Oxford Studies in Metaphyics, vol. 1*, pages 127–160. Oxford University Press, 2006.

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

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

J. Robert G. Williams. Normative Reference Magnets. *The Philosophical Review*, 2018.

J. Robert G. Williams. *The Nature of Representation*. Oxford University Press, forthcoming.

Timothy Williamson. *Vagueness*. Routledge, 1994.

Timothy Williamson. *Knowledge and Its Limits*. Oxford University Press, 2000.

Timothy Williamson. Must Do Better. In Patrick Greenough and Michael P. Lynch, editors, *Truth and Realism*, pages 177–187. Oxford University Press, 2006.

Ludwig Wittgenstein. *Philosophical Investigations*. G. E. M. Anscombe, trans. Blackwell, 1953.

Alex Worsnip. Disagreement as Interpersonal Incoherence. *Res Philosophica*, 96(2): 245–268, 2019.

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

Nick Zangwill. Moral Dependence. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics, vol. 3*, pages 109–127. 2008.