SEMANTIC ARGUMENTS AGAINST MORAL NATURALISM

by

Steven Coyne

B.A. Honours, University of Calgary, 2009

BSc., University of Calgary, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Philosophy)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2011

Abstract

This thesis investigates the prospects of a position in metaethics called moral naturalism. Moral naturalism can be summarized as two claims. First, moral naturalism is a form of moral realism, which states that there are true moral claims that hold irrespectively of a person's attitudes or beliefs. Second, moral naturalism claims that these moral claims are about properties that are part of the natural world.

The central challenge facing moral naturalism is to explain how these moral properties fit into the natural world. Are moral properties reducible to, or identical with, natural properties? If so, is there a semantic explanation for why moral properties are related to some natural properties, and not others? Two major arguments, the Open Question Argument and the Moral Twin Earth Argument, have suggested that such a semantic explanation is not possible, which would make moral naturalism an implausible position to hold. This thesis investigates the prospects for moral naturalism by assessing the success of these arguments.

The conclusions offered in this thesis are conservative. Both arguments turn out to depend on controversial, yet plausible, assumptions. In the case of the Open Question Argument, I argue that the success of the argument is sensitive to the form of moral naturalism under consideration; while it is fairly clear that it succeeds against reductive moral naturalism, it is less clear that it undermines non-reductive moral naturalism. It is clearer that the Moral Twin Earth Argument is successful, but it cannot categorically rule out every semantic explanation that the moral naturalist might advance.

Table of Contents

Abstract ..... **ii**

Table of Contents ..... **iii**

Acknowledgements ..... **v**

Chapter 1: Introduction **1**

1.1 Moral Naturalism vs. Moral Non Naturalism 1

1.2 The Metaphysics of Moral Naturalism 5

1.3 Explaining Moral Naturalism 10

Chapter 2: Against Analytic Moral Naturalism: The Open Question Argument **16**

2.1 Moore's Open Question Argument 16

2.2 Counterexamples and Open Questions 22

2.3 The Counterexample Argument for Moral Concepts 28

2.4 Are Moral Concepts Extensional? 29

2.5 Objections 33

Chapter 3: Against Synthetic Moral Naturalism: The Moral Twin Earth Argument **36**

3.1 The Moral Twin Earth Argument 36

3.2 Moral Twin Earth and Causal Regulation Semantics 42

3.3 Moral Twin Earth and Bare Descriptivist Semantics 49

3.4 The Moral Twin Earth Argument and Analytic Moral Naturalism 54

Chapter 4: Copp's Indirect Argument against Moral Twin Earth **56**

4.1 Copp's Indirect Argument 56

4.2 Responding to Copp's Indirect Argument 60

4.3 Generalizing Copp's Reply: The Metaphysics of Referential Intentions 64

4.4 Conclusion 69

Chapter 5: Conclusion 71

Bibliography 72

Acknowledgements

First of all, I owe an enormous debt of thanks to my supervisor, Dr. Matthew Bedke. In addition to his questioning and probing of the arguments in this thesis, he has been an excellent mentor more generally as I work towards becoming a philosopher.

I am also very obliged to the second reader of this thesis, Dr. David Silver. I have greatly appreciated his enthusiasm and attention to detail while working through this thesis.

This thesis benefited from conversations with Dr. Ori Simchen (who particularly influenced chapter 4,) and various graduate students around the department, such as Joshua Johnston, Catinca Mattice, and Jerry Viera. I presented Chapter 4 of this thesis at the UBC Graduate Colloquium and the Society for Exact Philosophy 2011 Meeting, and the audience questions from both talks were very insightful.

And special thanks to my family and Barbara, for moral support and proofreading on very short notice.

Chapter 1: Introduction

This introductory chapter aims to define moral naturalism, explain what is necessary to make moral naturalism a viable position in metaethics, and categorize various moral naturalist views into a few important groups.

1.1 Moral Naturalism vs. Moral Non Naturalism

Since moral naturalism will be discussed at length in this thesis, we should begin by asking what it is for something to be *natural.* Most philosophers endorse a roughly natural ontology, meaning that they reject the existence of things that do not belong to the natural world. In order to create everything in the world, all that was necessary was to put the natural things in place. But what is part of the natural world, and what is not?

We could take a first step towards answering this question by ostensively listing the things that we take to be natural. Chairs, particles and psychological states are all uncontroversially natural, whereas divine beings and unicorns are uncontroversially non-natural. Some kinds of things are a little more difficult to place into one category or the other; it is unclear whether sets, qualia, numbers, types, possibilities, or vague entities (like heaps) count as natural properties or entities.

The matter gets trickier when we attempt to find the features that natural things share in common. In particular, a necessary and sufficient criterion that distinguishes the natural from the non-natural is quite elusive. Consider the following attempt at a criterion: an entity is

natural if and only if it is the proper object of study by the empirical sciences.[1] To put this criterion in another way, Copp (2000) has suggested that "a property is natural if and only if any synthetic proposition about its instantiation that can be known, could only be known empirically." (39) But it's difficult to use this criterion to solve borderline cases – for example, mathematical entities like numbers are indispensible for science, but do not seem to be directly observable. Moreover, it is counter-intuitive to make what counts as natural depend upon the epistemic position of agents. Instead, one might suggest a *metaphysical* criterion for separating the natural from the non-natural. For example, one might think that natural properties are those that make a causal difference in the world. But metaphysical definitions of naturalness are also fraught with difficulties; as Shafer-Landau (2003) points out, surely the property of being self-identical is a natural property, even though it makes no causal difference in the world. Rather than taking a stance on this question, I will take for granted that there is an unproblematic set of entities that are natural, which in turn can be described using language that is unproblematically natural.[2]

An important problem in moral philosophy concerns the relationship between morality and the natural world. Do moral properties like rightness or goodness belong to the natural world, or are they something else?

There are appealing reasons in favour of the view that moral properties do not belong to the natural world. One reason is that they appear to possess an authority over our actions that

---

[1] This might have to be an ideal or complete form of natural science; for example, more empirical and theoretical work will need to be done to even understand the sense in which strings exist in string theory. But this ideal form of science would still need to be a recognizable natural science that makes empirical observations and compares hypotheses that explain these observations.

[2] However, this problem of criterion will come up again briefly when I discuss strongly non-reductive forms of moral naturalism.

natural properties or facts do not possess. Many philosophers hold that understanding a certain moral proposition – that an action is wrong, for example – is sufficient for giving the contemplator of that proposition a reason for or against performing that action, or even to motivate them not to perform that action, regardless of their beliefs or desires.[3] The epistemology of moral facts also appears to be significantly different from the epistemology of natural facts. For example, people appear to have incompatible positions on moral questions like abortion and the value of animal life, even when the non-moral facts relevant to the issue are agreed upon.[4] In contrast, our epistemic position appears to be much better with respect to natural facts. We have been very successful in resolving our empirical disagreements, and generally trust science to arbitrate these disputes. Another important difference between moral and natural epistemology is the fact that many philosophers think that we can discover or justify moral truths in an *a priori* fashion, without any further justification supplied by the external world.[5] This is vastly different from how we come to know empirical facts about the natural world, where we test hypotheses through observation and controlled experimentation. The most straightforward explanation for these epistemological differences is that moral facts are about a fundamentally different kind of entity than natural facts, if they are about any entity at all.

However difficult it might appear to be to resolve these differences, the non-naturalist faces an equally uncomfortable metaphysical burden: to account for a strange *sui generis* category

---

[3] Assuming that the belief is correct and the person is not mistaken. Also note that moral non-naturalism might also struggle to account for the motivational force of moral beliefs. If we grant that moral properties lack causal force in the world, it is hard to see how they could possibly motivate us.

[4] There is no shortage of literature in support of this claim. See Haidt (2001) for example.

[5] Intuitionists such as Shafer-Landau (2003), Huemer (2007) and Audi (2008) hold this view.

of properties that are not empirically investigable by science, and could themselves exert no causal force on non-moral affairs, even though they have empirical entities (such as actions or states of affairs) in their extensions. These would be properties above and beyond the properties of the natural world, and ontological parsimony suggests that we should avoid postulating unnecessary properties or entities if we can do without them.

There are many naturalist proposals explaining how to locate morality within the natural world. The simplest proposal consistent with naturalism is *irrealism* about morality, which is simply to deny that there are moral facts or properties. For example, one might hold that our moral evaluation reflects attitudes that we have towards certain practices, and that there is no sense in which attitudes can be true or false. However, the moral naturalism that will be discussed in this thesis is a form of *moral realism*. Moral realism can be encapsulated in three requirements:

1) *Cognitivism*. Moral realism holds that moral statements are descriptive and about the world; they express beliefs, and so can be true or false.[6]
2) *Objectivity*. Moral realism holds that moral statements are true or false regardless of the attitudes, beliefs, or psychological states of agents. This is not to say that the moral claims cannot involve these things; utilitarians, for example, hold that rightness depends on the consequences for the production of certain psychological states. However, the moral claims themselves are stance-independent.

[6] These different cognitivist claims are usually held simultaneously, but need not be. On an exotic theory of moral semantics, moral statements could be true/false, but not actually purport to describe anything. The moral naturalism that I will discuss, however, endorses all of these cognitivist claims.

3) *Non-triviality*. Even if we grant the prior two claims, it might be the case that all of the descriptive, stance-independent propositions are actually false. To be non-trivial, moral realism holds that there are at least some true moral propositions, which serve to attribute some properties that are actually instantiated in the world.

In the rest of this thesis, whenever I use the description *moral naturalism*, I mean this as shorthand for *realist moral naturalism,* which endorses all three of the above requirements, in addition to the naturalist requirement that nothing above and beyond the natural world is required for such statements to be true or false. [7]

1.2 The Metaphysics of Moral Naturalism

As stated in the last section, this thesis addresses the prospects of realist moral naturalism. The realist moral naturalist faces what Jackson (1998) terms the "location problem" for ethics. Where is morality *located* in the natural world? In this section, I will distinguish three positions regarding the way in which moral properties are natural, and in particular, how they relate to other natural properties.

Strong non-reductionism. On a strongly non-reductive version of moral naturalism, moral properties are natural properties because (and only because) they meet an epistemological or metaphysical criterion of naturalness, and they are not straightforwardly reducible to any other natural properties. While there might be true propositions detailing the relationships holding between moral and non-moral properties, there are no identities or equivalences holding between them. As a consequence, moral properties exert their own causal powers and are not at risk of redundancy in explanations of other natural phenomena such as actions.

[7] I am taking this criterion from Bedke's "Against Normative Naturalism", forthcoming in the *Australasian Journal of Philosophy.*

As Nicholas Sturgeon wrote of explaining the cowardly behaviour of the California pioneer Passed Midshipman Woodworth,

> But Woodworth not only failed to lead rescue parties into the mountains himself, where other rescuers were counting on him (leaving children to be picked up by him, for example), but had to be "shamed, threatened and bulled" even into organizing the efforts of others willing to take the risk .... DeVoto concludes: "Passed Midshipman Woodworth was just no damned good." (1942, p. 442) I cite this case partly because it so clearly has the structure of an inference to a reasonable explanation. One can think of competing explanations, but the evidence points against them.... (221)[8]

According to Sturgeon, the best explanations for certain actions irreducibly invoke moral properties, which themselves cannot be reduced to any other natural properties, even in particular instances.

Care must be taken to distinguish this position from Moorean non-naturalism. Moore thought that it would be impossible to supply a definition of 'good' because the property of goodness is not "complex":

> We may mean that a certain object, which we all of us know, is composed in a certain manner: that it is has four legs, a head, a heart, a liver, etc., etc., all of them arranged in definite relations to one another. It is in this sense that I deny good to be definable. (407)

[8] Arguably, this quotation might be taken as illustration of a weaker kind of non-reductive moral naturalism, where the moral facts still supervene on the non-moral facts.

To help explain this point, Moore draws an analogy between the definition of goodness and the definition of yellowness. Suppose that we can give a definition of the substantive to which the adjective 'yellow' will apply, perhaps objects whose atoms predominantly vibrate in certain ways in response to absorbing certain frequencies of light. This still does not provide a definition of yellowness, since this vibration of particles is not what we perceive, and Moore thinks that this perceptual quality is what we mean by 'yellow'. Therefore, yellow is a simple property that cannot be defined in terms of non-yellow things. Moore thinks that normative properties are also irreducible, perhaps because they have an irreducibly phenomenal character, as in the case of yellowness, but more likely because they exert different motivational force on people than that exerted by natural properties.

The difference between Moore's non-naturalism and strong non-reductive naturalism, then, is this: Moore thinks that moral properties are not investigable by empirical science, whereas a non-reductive naturalist thinks that they are empirically investigable. Strong non-reductive naturalism still counts as a form of moral naturalism because no ontology beyond that of natural world is required to make moral claims true or false, and we need to know Moore's views on moral epistemology in order to determine that he rejects this claim.

Weak non-reductionism. On a weak non-reductivist version of moral naturalism, moral properties are natural properties because (and only because) their *instances* bear appropriate relations to the instances of some other non-moral, natural properties. For example, while the rightness of Bob saving a cat from a fire can be identical with that action's selfless courage, the rightness of some other action (say Bob donating to UNICEF,) might be identical with its property of maximizing the happiness of the most people. Each individual instance of a moral property could have been, in principle, ascribed by *some* predicate couched only in

natural, non-moral language, but there is no purely natural predicate that could ascribe *all* instances of a moral property. Moreover, there is no need for any rules or laws that explain which instances of natural properties are identical to instances of moral properties.

I understand Shafer-Landau (2003) as providing an excellent example of such a view, in spite of his overall non-naturalist stance. He suggests that moral properties are *exhaustively constituted* by instantiations of natural properties.[9] What he has in mind by "exhaustively constituted" is not entirely clear, but a helpful hint is his comparison of moral properties to special science properties:

> There really are such things as being left-handed, crooked, rocky, Palaeozoic, molten, and deciduous, even if physics neither ratifies nor relies on such features, and even if there is no plausible path to the conclusion that such properties are identical to physical ones, disguised in more digestible vocabulary. (75)

We generally think that the truth of statements about special science properties, such as "Left-handed people are just as creative as right-handed people," requires no ontology beyond a naturalistically acceptable ontology, even though there are no hard and fast rules that allow us to ascribe them based on knowledge of their underlying physical basis. Yet every instance of a biological property is identical with some instance of a physical property, such as some atoms arranged in a particular way. So if moral properties are like special

[9] In a footnote, Shafer-Landau is explicitly neutral on the question of whether a given instance of a moral property is identical to the "natural facts that constitute it" (76), which seems to put him in the weak non-reductivist camp. However, Shafer-Landau thinks that the relevant difference amounts to the necessity of the token-token identities; if the identities are necessary, the non-naturalist should deny them, since this would prevent the properties from being realized in different ways.

science properties, no ontology above and beyond the ontology of the natural world is required for the truth of moral claims.

Reductionism. On a reductive version of moral naturalism, it must be possible to ascribe a moral property with a predicate that only uses naturalistically acceptable terms. This property may turn out to be very complex. For example, Boyd (1988) thinks that moral terms refer not to individual properties, but instead to *families* or *clusters* of properties:

> A family F of properties which are 'contingently clustered' in nature in the sense that they co-occur in an important number of cases... their co-occurrence is not, at least typically, a statistical artifact, but rather the result of what may be metaphorically (sometimes literally) described as a sort of homeostasis. (197)

These families of properties include the mechanisms that lead them to be clustered together. On Boyd's "homeostatic property cluster" account, moral goodness is a cluster of the many human goods, things important for satisfying needs, which tend to be mutually reinforcing.[10] It is likely that similar account could be drawn for other normative properties, like rightness or reasons.

Another form of weak reductionism holds that a moral property is a single disjunctive or context-sensitive natural property. For example, goodness might be identical with the disjunction of pleasure and intellectual accomplishment. In order to accommodate the wide variety of human goods, however, this property would have to be massively disjunctive, and

[10] See Rubin (2009), who argues that goodness cannot be a homeostatic property cluster. Rubin wonders: why do the various human goods have to be mutually reinforcing? Why should we rule out the possibility of goods that are necessarily isolated from other goods, in every possible circumstance?

sensitive to context. Depending upon how contextually sensitive the property is, a weak reductionism that advocates disjunctive properties may collapse into weak non-reductionism.[11]

1.3 Explaining Moral Naturalism

In the last section, I distinguished three positions that a moral naturalist might take regarding the relationship between the natural and the moral:

(1) Strong non-reductionism: moral properties are *sui generis* natural properties

(2) Weak non-reductionism: the instances of moral properties are identical to instances of natural properties, but no property-property identities hold.

(3) Reductionism: it is possible to ascribe each moral property with a predicate that only uses naturalistically acceptable expressions.

The truth of position (1) can be investigated empirically, but this does not seem very promising. Many of our paradigmatically natural properties, such as *being an atom*, are not *sui generis*, since they can be reduced to even simpler properties (such as being a certain number of neutrons, protons and elections arranged in a certain way), and might be understood as dispensible in scientific discourse. Since there seem to be very few genuinely *sui generis* natural properties – extension and mass are plausible candidates - it is unlikely that moral properties are among them. This position also seems to run contrary to the *supervenience* of the moral on the non-moral. Moral realists are generally wedded to the intuitive claim that there can be no difference in moral facts without a corresponding

[11] Many philosophers have also questioned whether disjunctive properties exist, or could have causal powers.

difference in the non-moral facts. If position (1) is correct, it seems as though moral facts can vary without any corresponding change in the non-moral facts. Positions (2) and (3), on the other hand, assert some sort of identity between moral properties and natural properties (or their instantiations). Consider some specific examples of how (2) and (3) might be correct:

(2*): the rightness of Bob saving a cat from a fire = the selfless motivation for that action

(3*): rightness = maximizing happiness.

Horgan and Timmons (1991) pose the following question: what is it about the world that makes facts like (2*) and (3*) true? One might think that we don't need to give any further explanation of facts like (2*) and (3*) – to put it another way, they are rock-bottom facts about the natural world. With respect to (3*), this would be for the following *sui generis* fact about 'right' to hold:

R: 'right' picks out the property of promoting the greatest happiness

This view is implausible for two reasons. First, we risk dramatically unhinging the meaning and reference of terms from the speakers who use them. If people can be *entirely* ignorant about the reference of our terms, this raises the worrisome possibility that we never have any idea what we are actually saying. Second, even if we can limit this ignorance so that it doesn't lead to a pervasive meaning-scepticism, we still expect that **R** is the sort of fact that need not be brute. We expect that the reference of our terms is exactly the kind of phenomenon that is amenable to further philosophical analysis. Or at the very least, it is the kind of fact that is sensitive to other non-semantic facts; it must be a synthetic fact that a word in some language refers to a certain property.

So we can safely reject the *sui generis* account of (2*) and (3*), and attempt to provide a semantic explanation of how they might be true. Philosophers think that the identities might be true or false in two different ways:

1) Moral-natural identities in (2)-(4) are *analytic*, so that only semantic facts about meaning explain why they are true.
2) Moral-natural identities in (2)-(4) are *synthetic*, so that in addition to semantic facts, facts about the world are also required to explain why they are true.

I will conclude this chapter by making a few remarks about each form of semantic explanation might be used to explain moral-natural identities such as (2*) or (3*).[12]

Analytic Moral Naturalism

An analytic truth, roughly, is a statement that is true strictly in virtue of the semantic meaning of the terms contained in it. For example, this is a paradigmatically analytic statement:

All bachelors are unmarried men[13]

By knowing the meaning of 'bachelor' and the meaning of 'unmarried man', one can know that all bachelors are unmarried men without any further investigation of the world. Sometimes this is put in terms of concepts: if a person is a competent user of the concepts *bachelor* and *unmarried man*, for any person to whom they apply the concept of bachelor, they should also apply the concept of unmarried man.

---

[12] For reasons given by Quine (1949), the distinction between analytic facts and synthetic facts is not perfectly clear, but a precise form of the distinction will not be crucial for my discussion.

[13] Many people have noted that this is not clearly a truth, let alone an analytic truth. For example, the Pope is an unmarried man, but we might hesitate to call him a bachelor. Counterexamples such as this statement hinge on whether our hesitance to apply the terms depends on the semantic content of the statements, or whether such hesitance stems from some other pragmatic factors.

Most philosophers, for reasons that will be discussed in chapter 2, reject such simple analytic equivalences. However, some philosophers have instead searched for more nuanced analytic views. For example, it is more plausible that a moral property could be analytically identical with a property that satisfies some description. Jackson (1998) has proposed a functional analysis of moral terms which he calls *analytical descriptivism*, which would vindicate type-type identities like that hold between moral and natural properties. On Jackson's account, the meanings of moral terms are given by the roles they play in a mature folk morality, which is a set of platitudes (such as 'pain is bad' or '"I cut, you choose" is a fair procedure') that he thinks a community will eventually converge upon. If we conjoin the platitudes regarding one particular role-property (say the 'good' role) and write them out as an open sentence, we can then look for the unique natural property that satisfies the sentence of good-platitudes. According to Jackson, we can then infer the identity of goodness and the natural property that satisfies the good-platitudes.[14]

In comparison with strictly analytic accounts of the truth of moral-natural identities, Jackson's account provides a constrained role for analyticity. According to Jackson, the biconditional containing the conceptual platitudes

x is good iff {*x* satisfies some set of mature folk platitudes about the good}

[14] There are two ways to interpret Jackson's functionalism regarding the reference of moral terms. We might think that the referent of 'good' is the property of satisfying some set of mature folk platitudes about the good. If so, the referent is known entirely analytically. On the other hand, we might think that the referent of 'good' is the natural property that satisfies the set of mature folk platitudes about the good. So we require both *a priori* and *a posteriori* facts to determine the referent of 'good'. In the latter case, it follows that 'good' is not a rigid designator.

is known analytically and *a priori*. On the other hand, the natural property that satisfies {*x* satisfies some role property given by a set of folk platitudes about the good} can vary from world to world and is known and identified *a posteriori*.

Synthetic Moral Naturalism

Another explanation for the truth of moral-natural identities such as (i)-(iii) is to appeal to some facts about the world, rather than only the meanings or semantics of the constitutive terms in the identity statement. Of course, all identities are true *partly* because of semantic constraints, because semantic constraints are necessary to determine which concepts are being expressed by an expression, or the referent of that expression. But according to a synthetic moral naturalist, these facts must be further complemented by facts about the world to determine whether the identities are true or false. For example, consider Kripke's causal-historical account of reference:

> A rough statement of a theory might be the following: An initial 'baptism' takes place. Here the object may be named by ostension, or the reference of the name may be fixed by a description. When the name is 'passed from link to link', the receiver of the name must, I think, intend when he learns to use it with the same reference as the man from whom he heard it. (96)

In order to determine the referent of an expression, we must know some facts about the property on the other end of the ostensive definition, or what property lies at the end of a chain of speakers. On a competing account advanced by Boyd (1988), which is more influential among moral naturalists, a term refers to a property (or cluster) if it is "causally regulated" by that property:

> *Roughly*, and for non-degenerate cases, a term *t* refers to a kind (property, relation, etc.) *k* just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term *t* will be approximately true of *k*. (195) We may think of the properties of *k* as regulating the use of *t*, and we may think of what is said using *t* as providing us with socially coordinated epistemic access to *k*; *t* refers to *k* (in nondegenerate cases) just in case the socially coordinated use of *t* provides significant epistemic access to *k*, and not to other kinds. (195)

Boyd's causal account and the Kripke-Putnam causal-historical account share some important features. Like Kripke's account, Boyd's account is externalist, for causal regulation permits the speaker to remain ignorant about the property that regulates their beliefs. And we might think that the causal chain connecting the baptism of a term to its use by other language users is one of Boyd's "causal mechanisms" that tend to bring about true beliefs regarding that term.[15] In both theories of reference, moral-natural identities are true in virtue of having both sides of the identity refer to the same property, and to know whether this is true, we must investigate the world (much like it took significant investigate to determine that the referent of 'water' was H_2O.)

[15] Geirsson (2005) raises the question of whether Boyd's causal regulation account implies that moral terms are rigid designators, and worries that if not, causal regulation will produce the wrong results in some of the cases that drove Kripke to a causal theory of reference for natural kinds like 'tiger'. I think it is reasonable, and in Boyd's spirit, to think of causal regulation as a reference-fixing proposal, instead of as a claim about the meaning of a term.

Chapter 2: Against Analytic Moral Naturalism: The Open Question Argument

In this chapter, I discuss the Open Question Argument (OQA), which is a major argument against analytic moral naturalism. In §2.1, I distinguish two distinct versions of the OQA that can be found in section 13 of Moore's *Principia Ethica*. In §2.2 I make some observations about the role of meaning in the OQA, and suggest a modification to the OQA (the "Extensional OQA") that provides stronger evidence against analytic moral naturalism. In §2.3, I argue that the extensional version of the OQA might be better understood as a push towards non-reductive moral naturalism, rather than as an argument ruling out moral naturalism altogether.

2.1 Moore's Open Question Argument

As discussed in chapter 1, moral naturalism asserts that moral properties are identical with natural properties, which is to say that moral properties *are* natural properties, even though they are presented under different (moral) names, like 'good' or 'right'. In order for this to be the case, the semantics of moral terms must be such that certain sentences expressing identities between the referents of moral and natural terms are true, for example that "maximizing happiness = rightness" is a true sentence. Moral naturalists are divided over how to explain the truth of such natural-moral identity statements, but one position on this question is *analytic moral naturalism*, which endorses one or both of these two closely related theses:

(i) The truth of a moral-natural identity statement can be explained by appealing strictly to the *meanings* of the moral and natural expressions in the statement.

(ii) The truth of a moral-natural identity statement can be explained by appealing strictly to the *a priori* aspects of moral and natural *concepts* expressed by the expressions in the statements.

It is not entirely clear whether one could hold one of these claims independently from the other, since meanings and concepts are thought to be closely related to one another. It might well be the case that our intuitions about the meanings of expressions provide us with direct access to the concepts expressed by those expressions. Regardless, I will assume in this chapter that it is relatively unproblematic to switch back and forth between meaning-talk and concept-talk.

In defence of their position, many analytic moral naturalists would make the following argument: if 'is right' is synonymous with 'promotes the greatest happiness', it follows that the properties ascribed by those predicates are identical, so long as the meaning of the expressions fix their reference.[16] The Open Question Argument (OQA), first raised by Sidgwick and Moore, denies the conclusion of this argument by denying the premise that moral expressions are synonymous with 'promotes the greatest expression', or any other expression couched solely in naturalistic vocabulary. It will be helpful to outline the OQA in premise-conclusion form:

(OQA-1) Analytic moral naturalism holds that moral-natural identities are true in virtue of meaning.

[16] I am using the word 'synonym' to indicate a relation holding between words with the same meaning.

(OQA-2) No moral expression is synonymous with any natural expression.

(OQA-C) Analytic moral naturalism is false.

It is clear that Moore, in section 13 of the Principia Ethica, intends to advance an argument along these lines, offering (OQA-C) as his conclusion. He writes:

> "...if I am right, then nobody can foist upon us such an axiom as that 'Pleasure is the only good' or that 'The good is the desired' on the pretense that this is 'the very meaning of the word'."

So long as we treat the quoted phrases in the above passage as describing moral-natural *identities* rather than simple biconditionals or predications, (OQA-1) and (OQA-C) seem to be faithful representations of Moore's intended conclusion in the OQA. The serious work to be done in this section is to explain how Moore justifies premise (OQA-2), and for that matter, to understand what Moore means by (OQA-2). As I have said, Moore's version of the OQA must show that for any 'X', where 'X' is an expression composed solely of non-moral expressions, 'X' and 'good' are not synonymous. Moore demonstrates this, roughly, by noting that we should have some awareness that 'X' and 'good' had the same meaning if they were indeed synonymous – and since we lack this awareness, it follows that the two expressions are not synonymous. All commentators on Moore agree on this much, but they often disagree about *how* Moore intended to coax our awareness of these non-synonymies out of us. Indeed, section 13 of the *Principia Ethica* contains two passages that might lead us to two different answers to this question. First is an introductory passage outlining the method that Moore intends to employ against moral naturalism:

> [Passage 1]: The hypothesis that disagreement about the meaning of good is disagreement with regard to the correct analysis of a given whole, may be most plainly seen to be incorrect by consideration of the fact that, whatever definition be offered, it may always be asked, with significance, of the complex so defined, whether it is itself good. (411)

This first passage argues that, for any proposed analysis 'X' of 'good' into non-moral vocabulary, it "may always be asked, with significance, of ['X'] so defined, *whether it is itself good*".[17] The argument presented here can be framed as follows:

1) For any proposed definition 'X' of 'good', it can be asked with significance of the complex 'X' whether it is good.
2) If it can be asked with significance of 'X' whether it is good, 'X' cannot be the correct analysis of 'goodness'.
3) 'X' cannot be the correct analysis of 'goodness'.

In its present form, this argument is puzzling. What does it mean to ask, *of a complex*, whether it is good? Moreover, what kind of thing is a complex? Is it a property, some sort of semantic device, or a set of entities that have a property? In all three cases, it seems odd to predicate goodness of the thing in question. As Darwall, Gibbard and Railton (1992) have noted, it is not automatically guaranteed that a given complex "X" itself has the property of X-ness. Is oddness itself odd? Is non-self-identity non-self-identical, automatically? And even if they have these properties, do they have them trivially and *a priori*? These questions

[17] As Feldman (2005) notes, Moore often does not distinguish between the mention and use of a term – especially regarding a complex 'X'. So the referent of 'itself' in 'is itself good' could be any number of things.

of self-predication are difficult to resolve, and suggest that we should look to frame the argument in a different way.

In the following passage, Moore suggests that he will illustrate how to carry out this method, but ultimately differs from the argument of the first passage by correcting the problem mentioned in the last paragraph. He substitutes 'X' with the expression 'what we desire to desire' and concludes that 'what we desire to desire' is not synonymous with 'good':

> [Passage 2]: But, if we carry the investigation further, and ask ourselves 'Is it good to desire to desire A?' it is apparent, on a little reflection, that this question is itself as intelligible, as the original question 'Is A good?' – that we are in fact, now asking for exactly the same information about the desire to desire A, for which we formerly asked with regard to A itself. But it is also apparent that the meaning of this second question cannot be correctly analyzed into 'Is the desire to desire A one of the things which we desire to desire?': we have not before our minds anything so complicated as the question 'Do we desire to desire to desire to desire A?' (411)

The argument of the second passage differs from the first in two important respects. First, rather than addressing the goodness of the complex 'X', it addresses whether some object A is in the extension of the complex. Moore is not investigating whether *what we desire to desire* is good, but rather, whether *the desire to desire A,* presumably for some arbitrary A, is good. Second, Feldman (2005) notes that this passage compares *pairs* of declarative sentences or questions, and asks whether the individual *sentences* or *questions* are synonymous with one another. Consider the following pair of questions:

(Q1): Is it good to desire to desire A?

(Q2): Do we desire to desire to desire to desire A?

In the second passage, Moore notes that when we contemplate (Q1), "we do not have before our minds anything so complicated as [Q2]," and presumably it follows from this that (Q1) and (Q2) differ in meaning. To arrive at the further conclusion that 'good' does not mean 'what we desire to desire', Feldman suggests that Moore implicitly relies on a principle of compositionality, which requires that the meaning of whole sentences be fixed by the meanings of their individual parts. Given that (Q1) and (Q2) apparently have different meanings, it must follow from compositionality that this difference is a result of the different meanings of the only dissimilar syntactic parts of the sentences, which are the predicates 'good' and 'desire to desire X'. The argument in the second passage can be framed as follows:

1) (Q1) and (Q2) do not place the same notion before the mind.
2) (Q1) and (Q2) do not have the same meaning.
3) The only syntactic difference between (Q1) and (Q2) is the substitution of the expression 'good' with the expression 'what we desire to desire'.[18]
4) The expressions 'good' and 'what we desire to desire' are not synonymous.

It is noteworthy that the argument in the second passage does not rely on the significance or openness of any question at all. If there is any role in the argument for the significance of certain questions, it is to serve as further support for the first premise. Moreover, this is the pattern of argument used in the remainder of the section. The two arguments that

[18] Obviously this is only true if we take some liberalities with the various connective words, because English treats the predicates slightly differently in terms of grammatical syntax.

immediately follow this passage – the passage beginning with 'Moreover', and the argument against defining 'good' as pleasure – also argue for premise 1 and arrive at the conclusion in the same manner.

In the rest of this chapter, I argue against the Open Question Argument (OQA), which is often taken to refute all forms of analytic moral naturalism. First, I observe that the OQA, in both passage 1 and passage 2 versions, falls short of the usual standards of demonstrating the irreducibility of one concept in terms of another. Second, I suggest a close relative of the OQA that meets these standards. Third, I argue that this strengthened version of the OQA does not undermine all forms of moral naturalism, but instead gives us insight into the peculiarly non-reductive structure of moral concepts.

2.2 Counterexamples and Open Questions

Many philosophers believe that certain important philosophical concepts, such as knowledge or motivating reasons, can be analyzed in terms of simpler, less controversial concepts. For example, Davidson (1963) argued that *motivating reasons* should be analyzed in terms of the relevant action being caused in an appropriate way by a belief and a desire. If we grant that there are at least some non-trivial analyses of philosophical concepts, the common epistemic standard for rejecting an analysis is to provide a *counterexample* against that proposed analysis. At the beginning of Plato's Republic, Socrates provides such a counterexample against Cephalus' analysis of justice as speaking the truth and paying your debts:

> Well said, Cephalus, I replied; but as concerning justice, what is it? --to speak the truth and to pay your debts --no more than this? And even to this are there not

> exceptions? Suppose that a friend when in his right mind has deposited arms with me and he asks for them when he is not in his right mind, ought I to give them back to him?

It is surprising, then, that many contemporary philosophers accept a lower standard of evidence than Socrates when rejecting analyses of moral concepts. The most influential global rejection of such analyses is provided by Moore's Open Question Argument (OQA), described in the last section.

Much of the critical attention on the OQA has focused on the kinds of moral-natural identities that escape it. Synthetic identities, such as the identity between water and H_2O, escape the reach of the argument because they require further *a posteriori* evidence for their justification – it is no wonder they are open to semantically competent speakers.[19] This paper attacks the OQA from a different angle. I will show that the epistemic standards of the OQA are lower than those demanded elsewhere in philosophical analysis. No one believed that the mere openness of the question "Granting that he is acting for a (motivating) reason, is his action caused in an appropriate way by a belief and desire?" was sufficient for undermining Davidson's analysis of motivating reasons. Nor did anyone believe that the openness of the question "Granted that it is justified true belief, is it knowledge?" refuted the analysis of knowledge as justified true belief. It took the difficult work of finding counterexamples (where they could be found) to convince people that these analyses were unsuccessful. I will argue that there is no good reason for treating the moral case differently from the usual philosophical case, and therefore that the OQA should be revised to meet these standards.

---

[19] It may be argued that an extended form of the Open Question Argument, where the semantically competent speaker is aware of all the non-moral facts, is successful against these synthetic identities. See Bedke (2011).

Let's begin by distinguishing between two methods for demonstrating the irreducibility of a concept X in terms of concept Y:

1) *The Counterexample Method*, which relies on evidence that the extensions of 'X' and 'Y' are different.[20]
2) *The Open Question Method*, which relies on evidence that the question "Granted that it is X, is it Y?" feels open.

I have said that the Counterexample Method provides stronger justification against a conceptual analysis than that provided by the Open Question Method, but I have not yet given an argument for that claim. It will be helpful to think of *concepthood* as something which can also be analyzed. If so, there are conceptual prerequisites to whether a given pair of expressions (X,Y) express the same concept. Even if we cannot provide a total analysis (that is, without remainder) of this relation of same-concept in terms of simpler concepts, we can likely find some simpler concept C that is *partly* constitutive of it. If so, our knowledge about the reducibility of concept X in terms of concept Y should be a consequence of our knowledge of whether the pair (X,Y) falls under the extension of C.

There are many plausible candidates for C, such as its application conditions or intension, but one is uncontroversial: its extension. If two expressions express the same concept, they must

[20] This evidence is defeasible. Sometimes a particularly explanatory theory about the conditions for applying a concept might lead us to reject some intuitions about outliers in its extension. For example, my intuition about the trolley cases, while strong, might be insufficient to lead me to reject an analysis of rightness in terms of maximizing happiness.

have the same extension.[21] By demonstrating that the two expressions do not have the same extension, the Counterexample Method shows that the two concepts must be different.

Does the Open Question Method appeal to any conceptual constituent of concepthood to illustrate the irreducibility of one concept in terms of another? Surely the openness of certain questions isn't itself a *conceptual* prerequisite to the non-identity of concepts X and Y in the same way that their differing extensions are. So how is it that the openness of a question is supposed to give us information about the relation between two concepts? In particular, is the openness of the question a consequence of some direct cognitive access to the structure of these concepts, or is there a more indirect conduit for this access?

We can answer this question by investigating our mental goings-on as we contemplate those open or closed questions. These goings-on are most transparent as we contemplate a question that is widely thought *not* to be open:

> "Granted that he is a bachelor, is he an unmarried man?"

We hum and haw, searching for a meaningful difference between a bachelor and an unmarried man, and when we find such a difference, it generally seems to take the form of a counterexample.[22] This suggests that it isn't the difference in conceptual meaning *itself* that is proximally responsible for the openness of the question. Rather, the difference in concept produces a difference in extension, which in turn produces a feeling of openness. So we can say that the Counterexample Method strictly dominates the Open Question Method in terms

---

[21] At least, barring any contextual or indexical effects.

[22] Indeed, some people have thought that the pope counts as a counterexample to the analysis of 'bachelor' as 'unmarried man'.

of the justification provided against the irreducibility of one concept in terms of another. Since there are more steps in the mechanism, there are more opportunities to go awry.

The case of pejoratives shows how we are often derailed by misleading feelings of openness. It is tempting to find the following question open:

> "Granted that he is an Irishman, is he a Mick?"

But this would be a mistake. Irishman and Mick are concepts of the same thing, and they are extensionally equivalent. The feeling of openness when contemplating this question is driven by non-semantic attitudes that are carried along with the semantic, belief-type attitudes; in addition to expressing a belief, they also express disapproval or disdain of their objects.

It would be ideal if we could further show that the irreducibilities *must* be explained by such a difference in extension of the irreducible concepts. If so, we could say that the failure for an ideally conceptually competent person to find a difference in extension between 'good' and 'X' means that their feeling about openness must be incorrect. This connection between conceptual irreducibility and difference in extension is more plausible when we amend our argument to consider the extension of expressions not only at this world, but at *all* possible worlds:

> If the extensions of 'X' and 'Y' are the same at every epistemically possible world, then the concepts expressed by "X" and "Y" are identical.

I take it that the main difficulty with the reductive claim above is the following: one might think that even though the extensions of 'X' and 'Y' are the same, they could name different properties because the property named by 'Y' has some feature that the property named by

'X' does not possess, or vice-versa. Indeed, this was clearly a worry that drove Moore to moral non-naturalism. Moore gives an extended discussion of the irreducibility of 'yellow':

> There is, therefore, no intrinsic difficulty in the contention that 'good' denotes a simple and indefinable quality... Consider yellow, for example. We may try to define it, by describing its physical equivalent; we may state what kind of light-vibrations must stimulate the normal eye, in order that we may perceive it. But a moment's reflection is sufficient to shew that those light-vibrations are not themselves what we mean by yellow... (10)

I understand this worry in the case of synthetic conceptual identities, such as how yellow cannot merely be defined as the kind of light-vibrations that stimulate the normal eye. The former has a feature that the latter does not, namely the conscious character of the experience.[23] But I have trouble extending this worry to the case of analytic, *a priori*, conceptual identities. Consider the following example: suppose that "justified true belief + X" had the same extension as "knowledge", so that every instance of knowledge was also an instance of JTB+X, at every epistemically possible world. Could we imagine a scenario in which the property of knowledge had some feature that JTB+X did not? This seems implausible if they really are extensionally equivalent. Likewise, compare the properties of being three-sided and three-vertexed. Could we imagine some property holding of one, but not the other?[24] I do not believe so. Every example that comes to mind, such as "composed of

[23] At least, Moore's argument is furthered if we grant him this controversial point in the philosophy of mind.

[24] Jackson (1997) gives an argument that necessarily co-extensive properties are identical.

three straight lines" or "has internal angles that sum to 180 degrees," must be predicated of both figures if predicated at all.

To sum up, there is a good reason for requiring the additional justification provided by the Extensional Open Question Argument: two expressions must differ in extension, at least at some possible world, in order to express different concepts.

2.3 The Counterexample Argument for Moral Concepts

The argument from the previous section suggests that a semantic argument against analytic moral naturalism should steer clear of open questions about analyses, and instead take the following form:

Counterexample Argument (for moral concepts):

For all analyses 'X' of 'good' couched only in natural (non-moral) vocabulary,

(1) For some *x* that is known by the speaker to have the property described by 'X', if the question "Is *x* good?" would **seem open** at some possible world to a semantically competent speaker, then 'X' and 'good' express different concepts.

(2) For some *x* that is known by the speaker to have the property described by 'X', the sentence "Is *x* good?" would **seem open** at some possible world to a semantically competent speaker.

(3) 'X' and 'good' express different concepts.

(4) The concept goodness cannot be analyzed in terms of concept X.

Consider an attempt to analyze 'good' as '*pleasant'*, and take the counterexample *x* to be *feasting on hors d'oeuvres*. I agree that feasting on hors d'oeuvres is pleasant, but I have doubts about whether it is good. At the very least, the question seems open to me. So it follows that goodness cannot be analyzed as pleasure. Moreover, it is tempting to grant that this argument extends to any choice of X that the moral naturalist might propose as an analysis of 'good'.

However, as this argument has been set up, it is not entirely parallel to the Counterexample Method. I take it that our mental goings-on in the moral case are different than when contemplating the Gettier cases. On the one hand, the feeling of openness is weaker than that of the Gettier case.[25] On the other hand, there is a sense of openness *whenever* I ask if some entity or object *x* is good. In the next section, I will offer an explanation for this difference.

2.4 Are Moral Concepts Extensional?

In light of the argument in the last section, it is tempting to grant that one can coherently ask of any *x* whether or not it is good. Now this does not by itself imply that *x* does not fall under the extension of good, but it does raise a worry – why are moral concepts so different from ordinary concepts, of which we have clear senses of which entities fall under them? One explanation, characteristic of non-cognitivism, is that moral concepts lack extensions. If so, the apparent extensionality of moral concepts can be chalked up to other pragmatic factors that have nothing to do with the semantic content of the concept. A closely related view,

[25] We might express this as the "modal force" that attends the contemplation of the question, as described by Bedke (2008). In the case of knowledge, it seems impossible that the Gettier case could count as knowledge, in some sense of the word impossible.

proposed by Bedke (unpublished), is that moral concepts are extensional, but it is semantically optional whether any entity falls under them. In this section, I will argue that the data from the Counterexample Argument need not indicate the *non-extensionality* or *semantic optionality* of moral concepts, but could instead support the *underdetermination* of moral concepts, which amounts to endorsing a strongly non-reductive form of moral naturalism.

The feeling that a Gettier case does not constitute knowledge is immediate and cannot be further justified. In contrast, the question "Is *feasting on hors d'oeurves* good?" feels very open, and seems like I could give considerations that weigh in favour of either answer. On the one hand, I might reason that feasting on hors d'oeurves is a pleasant experience. On the other hand, feasting on hors d'oeurves is sometimes overindulgent, which leads me to doubt that it falls under the extension of 'good'. This can be fixed, however. When I amend the question to "Is *feasting on hors d'oeurves without overindulgence* good?", it feels *less* open. We might express this as a difference in the credence we have towards an entity described with each of the proposed descriptions falling under the extension of 'good':

> Credence(that *feasting on hors d'oeurves without overindulgence* falls under the extension of 'good') > Credence(that *feasting on d'oeurves* falls under the extension of 'good')

Our credence in the proposed analysis can increase as we provide an increasingly detailed description of the entity under consideration as falling under the extension of 'good'. So for certain choices of additional descriptive content Y,

> Credence('x+Y is good') > Credence('x is good')

To see another example of how this process is carried out, consider whether pleasure counts as good. My credence that pleasure is good is initially is not very high; the question feels very open to me. But if I specify the kind of pleasure, or explicitly exclude certain pleasurable but clearly bad things (such as the pleasure that a sadist experiences when inflicting pain on somebody else,) my credence in the extension increases:

> Credence(pleasure not obtained through masochistic means is good) >
>
> Credence(pleasure is good)

Now we can contrast this process with the analysis of a concept that is genuinely semantically optional across a range of cases, such as whether a glider counts as an airplane.[26] A powered aircraft like a Cessna is clearly an airplane, and a helicopter is clearly not an airplane. But there are a range of cases, such as gliders, whose airplane-hood seems to be semantically optional. I can give considerations that seem to count in favour of a glider being an airplane (it flies and has wings,) but can also give considerations that count against it being an airplane (it is unpowered).[27]

We can now highlight an important difference between an expression with a partly semantically optional extension (like 'airplane') and an expression with a "convergent" extension (like 'good'). Imagine that we complement the description 'glider' with some additional descriptive content Y, just as we did when contemplating whether feasting on hors d'oeuvres is good:

---

[26] This example can be found in Bedke (unpublished).

[27] There is an interesting distinction here. Some concepts might be semantically optional for no further reason, but in this case, part of the concept's domain is semantically optional because competing reasons apply.

Credence(that a glider is an airplane) = Credence(that a Y glider is an airplane)

It will matter that Y does not consist of something we already know about a glider. We already know that a glider is unpowered, and that it is capable of flight – that is part of what it is to be a glider. However, not all gliders are red, so take Y to be 'red'. I don't have any stronger intuitions about whether a red glider is an airplane than whether an ordinary glider is an airplane. I suspect that this can be extended to any choice of additional descriptive content Y; as much as you expand upon the description of the glider (so long your choice of description isn't self-contradictory, such as "mechanically-powered glider",) the credence of whether it counts as an airplane should not change. I suspect that this shows that the semantics of 'good' and 'airplane' are different in an important way. To sum up, we might endorse the following position about the semantics of 'good':

> *Convergent Moral Naturalism:* For some choices of 'Y', if we add 'Y' to our description of an entity 'x', it is increasingly part of the extension of 'good'.

Though I suspect that some descriptions of entities are sufficient to remove all doubt about whether the entity falls under the extension of 'good', some might argue that this process will not converge for any description. For example, if I ask whether pushing people into puddles is wrong, there might be no set of further conditions that guarantees that this is wrong. Nonetheless, I maintain that adding further conditions (I wasn't coerced, etc) increases the wrongness of this action.[28] A helpful model of this form of vagueness is given by multi-

[28] The critic of non-reductive analytic moral naturalism might observe that 'good' and 'right' are likely to be indeterminate if they are derivative on whether a given feature of a situation *counts as a reason* for acting/not acting, so that even though there are truths about whether a feature of a situation counts for or against an action, there are no general rules regarding whether they make a given action right or wrong. If so, it is no surprise (semantically) that good or right are indeterminate, but we shouldn't expect this to hold of reasons. For

valued logic. Instead of a concept partitioning the domain of entities into its extension, anti-extension, and optional cases, we might assign a number to each entity that signifies "how much" it falls within the extension of the concept. If this model is correct, the fact that no entity falls completely under a concept will be completely innocuous; perhaps the best we can do is to say that pushing people into puddles is mostly wrong.

2.5 Objections

In this section I will respond to three objections that might be levied against Convergent Moral Naturalism.

The Non-Naturalist's Objection. When we focused on an *analysis* X of goodness, it was clear that the analysis X was a natural concept. But Convergent Moral Naturalism only supplies us with a concept whose *extension* (if it has even has one, in the proper sense of the word) consists of natural entities, and this is fully consistent with Moorean moral non-naturalism, which holds that goodness is an irreducible and non-natural concept whose extension consists entirely of natural entities.

Reply: One might argue for the following condition for a concept to count as natural:

> All and only natural features are responsible for determining whether (and how much) an entity falls under the extension of 'good'.

---

example, surely we don't need to provide additional descriptive content to determine whether "the fact that it causes arbitrary pain to someone" counts as a reason against performing that action. This seems to wed the non-reductive analytic moral naturalist to a very strong form of moral particularism.

If so, Convergent Moral Naturalism leads us to a naturalistically acceptable concept of goodness. Now, it is tempting to confuse this condition with supervenience, which holds that there can be no moral difference without a natural difference. And a non-naturalist such as Moore would be willing to endorse a supervenience thesis about moral properties. However, this condition is stronger than supervenience. We have claimed that there are analytic, semantic reasons for why the moral differences supervene on the natural differences, which is why the natural features are "responsible" for the extension of the concept.

The Metaphysician's Objection. In order to count as a naturalist theory of goodness, the concept *good* must be a concept *of a natural property*. But the natural property posited by Convergent Moral Naturalism is semantically very strange, and unlikely to exist in the real world.

Reply: To bolster the case it is reassuring to note that there are perfectly respectable natural properties that are semantically similar. If we were to attempt to find an analysis of kind expressions like "game" or "health", I suspect that we would fail. There is no description of a person that is sufficient to guarantee that they are healthy, and nor does knowing that they are healthy seem to entail anything (other than the fact that they are alive!) And apart from naming a few games (backgammon and chess are games,) it would be very hard to find necessary truths about games themselves. As Wittgenstein noted, there is a family resemblance holding between different kinds of games, but no umbrella property unifying them all. Some games involve winning or losing, but not all. Some games involve boards and pieces, but not all. In spite of this, the properties ascribed by "health" or "game" seem to be

unproblematically natural, and we routinely cite them in naturalistically acceptable explanations.

The Disagreement Objection. Consider a disagreement between Culture A and Culture B over a moral issue, for example, whether vegetarianism is morally required. Culture A and Culture B have different beliefs about the extension of the concept *morally required*, and assuming that they fully semantically competent and aware of all the relevant non-moral facts, this seems to suggest that they are using different concepts. Yet it is tempting to think that Culture A and Culture B are both semantically competent users of the concept.

Reply: The standard view of concepts, with a strict extension and anti-extension, suffers badly from this objection. If the two cultures disagree even slightly over the extension of the concept, they cannot be understood as using the same concept. But Convergent Moral Naturalism offers a more plausible reply: by attempting to fit terms like 'morally permissible' and 'right' to a concept that is actually rather vague, it is inevitable that differences will arise in their application.

Chapter 3: Against Synthetic Moral Naturalism: The Moral Twin Earth Argument

In the last section, I argued that the strongest form of the Open Question Argument escapes several objections, but may not undermine all forms of analytic moral naturalism. But I have not yet considered the possibility that there might be a synthetic relation grounding the identity between moral properties and natural properties. In this section, I will assess a recent and influential objection against synthetic moral naturalism called the Moral Twin Earth Argument, which was advanced by Horgan and Timmons (1991) and has since inspired a number of responses by moral naturalists.

3.1 The Moral Twin Earth Argument

R.M. Hare gave an early version of the Moral Twin Earth argument that only applies straightforwardly against analytic moral naturalism, but it is helpful for the sake of introducing the argument against synthetic moral naturalism. His version of the argument invites us to consider two distinct communities – one consisting of missionaries, the other consisting of cannibals. These communities are to be imagined as having very different views about morality, and also speaking different (but largely intertranslatable) languages. Hare then entertains the hypothesis that missionaries and cannibals intend to use their moral terms (like 'right' or 'wrong') to attribute properties. Suppose that missionaries and cannibals use the word 'right', which is "the most general adjective of commendation" (148) in both of their languages, in the following ways:

By ‘right’, missionaries mean ‘consistent with the Ten Commandments’ (1)

By ‘right’, cannibals mean ‘productive of the most scalps’ (2)

If the hypothesis that moral expressions serve to attribute properties is correct, Hare notes that since missionaries and cannibals *mean* to attribute different properties with the word ‘right’, speakers from each community talk past one another when disagreeing over whether to attribute moral properties to a common action. To see how this failure to capture the correct sense of their disagreement might come about, consider the following exchange between a missionary and cannibal:

Missionary: “You really should stop taking scalps in battle. It’s wrong.”

Cannibal: “Hrmph, I beg to differ. Taking scalps in battle is *exactly* the right thing to do.”

How should we understand this exchange?[29] Given (1), the missionary intends to use the term ‘wrong’ to attribute the property of *being inconsistent with the Ten Commandments*, so it follows that the sentence expressed by the missionary is plausibly true, depending on how one interprets the Ten Commandments. On the other hand, given (2), the cannibal intends to use the term ‘right’ to attribute the property of *being productive of the maximal number of scalps* and it follows that the cannibal has also expressed a true sentence – indeed, an almost analytically true sentence.

However, when we look at the sentences of the missionary and cannibal side by side, this data contradicts our intuition that they really are having a disagreement – surely the two

[29] In answering this question, we should further assume that the missionary and the cannibal have the same understanding of the non-moral terms in their sentences, so that there is no disagreement on what it is to take a scalp.

sentences aren't consistent with one another! Since we think that missionaries and cannibals are not talking past one another, moral language must serve some further function than to attribute some natural property to an action. Hare thinks that our willingness to translate 'good' between different languages, and consequently our intuitions about the disagreement between speakers from different linguistic communities, are best explained by 'right' primarily serving to express an evaluative attitude such as commendation.[30] So if there are any truths to be explained about the relationship between moral properties and natural properties, the connection cannot be analytic.

In response, the moral naturalist might argue there is a fact of the matter whether the missionary or the cannibal (or neither) are using normative terms like 'good' or 'right' correctly. Hare's argument tacitly assumes that the meanings of the words employed by missionaries and cannibals, encapsulated in (1) and (2), are known to the speakers and reflect their moral beliefs.[31] If this is the case, there is no clear sense that they can use moral terms erroneously, for the meaning of the terms is exhausted by what speakers take these meanings to be. There are a number of internalist metasemantic positions that guarantee this impossibility of error, and a well-known example of such a position is the descriptivism often attributed to Russell.[32] We might characterize descriptivism in the following way:

---

[30] Hare's example actually focuses on 'good', but to maintain consistency with the contemporary Moral Twin Earth argument I will use the example of 'right'. Occasionally I will switch back and forth between the examples for the ease of incorporating quotations from other philosophers.

[31] It need not reflect *all* of their moral beliefs, of course. A person can still be in error about particular cases, etc.

[32] Note that descriptivism permits that a speaker to be ignorant about which property actually satisfies the description they associate with a given term.

> **$R_{descriptivism}$**: a term *t* in a language L refers to a given property *p* if and only if speakers of L associate a description with *t* that *p* uniquely satisfies.

In particular, if a speaker or whole linguistic community already associates a certain natural property (under some description in their language) with a moral term, it follows that the moral term refers to that natural property. And this appears to be the metasemantic picture that Hare appears to be attacking: the hypothetical missionaries are aware that the missionary word 'right' means 'consistent with the ten commandments', just as the cannibals are aware of the meaning of the cannibal word 'right'.

However, these internalist theories of reference compete with a number of externalist theories of reference. These theories were advanced in the 1970s and 1980s by "New Wave" philosophers of language such as Kripke, Putnam, and Burge, who noted that we can be unaware of which property or entity is the referent of our terms, or even of the metasemantic facts which determine the referents of our words – so we might not be in a position to recognize the referents of our terms even under ideal epistemic conditions. On these accounts, social facts and facts about the natural world can play an important role in determining the reference of terms. For example, a baby is baptized "John", and the name gets passed along a complicated causal chain, but at the end of this causal chain a speaker can use the name and successfully refer to the original baby John, in spite of the fact the speaker may have no correct beliefs about John, or knowledge of the initial baptism or causal chain proceeding from it.

This externalist picture of language is applicable to several categories of words, but its applicability to natural kind terms is supported by a number of thought experiments

involving proper names and natural kind terms. A prominent example of such thought experiment, Putnam's Twin Earth, asks us to imagine a replica of Earth called Twin Earth. On Twin Earth, people act in the same ways, have the same mental states, have languages composed of lexicographically identical words, and the two worlds are almost entirely identical. There is only one difference between Earth and Twin Earth: all of the H_2O on Twin Earth has been replaced by a substance XYZ that behaves in exactly the same way as H_2O, but has a different molecular composition. Putnam claims that even though the mental states of speakers on Earth and Twin Earth are otherwise identical, the Earth word 'water' will have a different meaning than the Twin Earth word 'water'. This intuition is elicited by the observation that people on Earth and Twin Earth will be talking past one another when using their otherwise lexicographically identical word 'water'.

If an externalist theory of reference were correct, the moral naturalist could respond to Hare's argument by pointing out that the situation involving missionaries and cannibals is constructed somewhat artificially. Even though it is possible that the meaning of 'right' might come apart for the two communities, it is also possible that the term 'right' has the same meaning in the missionary language as it does in the cannibal language, despite the unawareness of missionaries or cannibals of this fact. Instead, some set of external facts could ensure that their moral terms refer to the same property. In Boyd's example, despite their mistaken beliefs about 'right', perhaps their uses of the word 'right' are regulated by the same property.

Horgan and Timmons constructed a version of Hare's missionary and cannibal argument that addresses versions of synthetic moral naturalism motivated by these "New Wave" semantic

theories. Their objection has come to be referred to as the Moral Twin Earth argument or the Moral Twin Earth objection.

The objection proceeds in the following way. Consider any reference relation **R** that makes a natural property *k* the referent of a moral term *t*; it may help to imagine that a causal-historical or descriptivist theory of reference constitutes **R**. Horgan and Timmons claim that one of the following conditions applies to any account of **R:**

(1) **R** fails to pick out a determinate property for a moral term *t*, because there are too many natural properties that could serve as the referent for *t*.

or

(2) **R** allows for *k* to differ between two speakers from two different linguistic communities[33]

It is clear that (1) would be a major problem for an account of the metasemantics of moral terms, as the metasemantics would be unable to provide the explanation of the moral-natural identity that was the whole point of providing a semantic account of the referent of a moral expression. We have already seen from Hare's version of the argument that (2) is also a problem for any metasemantic account of moral terms. If any synthetic theory of reference is stymied by one of these two conditions, it follows that moral terms cannot be typical property-ascribing predicates, and so the moral-natural identities posited by moral naturalism cannot be given a synthetic grounding.

---

[33] In discussing this horn of the dilemma, Horgan and Timmons (2000) actually distinguish between two kinds of relativism: chauvinistic relativism and standard relativism. According to chauvinistic relativism, only one of the two communities has the correct concept of 'right'. But they claim that this possibility is also undesirable, and I will grant this without further criticism.

At this point, a critic might lodge an obvious objection: Horgan and Timmons have not shown that *every* naturalistic metasemantic relation **R** must meet one of conditions (1) or (2). Why can't there be a relation **R** that picks out a determinate referent property for *t*, yet also picks out the same property from linguistic community to linguistic community? We might have the lingering feeling that a more sophisticated semantic account can avoid both of conditions (1) and (2). While I will not provide a full generalization of the Moral Twin Earth argument to all possible accounts of **R**[34], in the next two sections I will show how the argument applies against

a) a characteristic example of externalist metasemantics - the "causal regulation" semantics proposed by Boyd (1988), and
b) a representative example of an internalist metasemantics that is more nuanced than the account assumed in Hare's discussion.

3.2 Moral Twin Earth and Causal Regulation Semantics

In the course of defending his version of moral naturalism, Boyd characterizes the reference-fixing relation **R** in the following way:

Roughly, and for non-degenerate cases, a term *t* refers to a kind (property, relation, etc.) *k* just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term t will be approximately true of k.

[34] In order to do this, one must find a useful way to categorize the numerous accounts of reference. One property that might be useful in categorizing them is the internalist/externalist distinction. Another possible division is to focus on how different referential accounts have words "hook up" to the world; philosophers such as Kripke appear to have held that ostension and description exhaust the ways in which a word might describe something in the world.

> We may think of the property *k* as regulating the use of *t*, and we may think of what is said using *t* as providing us with socially coordinated epistemic access to *k*; *t* refers to *k* (in nondegenerate cases) just in case the socially coordinated use of *t* provides significant epistemic access to *k*, and not to other kinds. (195)

To see Boyd's metasemantic theory at work, consider how it maps natural kind terms such as 'water' to natural properties. The word 'water' was coined to describe a liquid substance that boils at a certain temperature, is a good solvent, and so on. Even before people understood molecular chemistry well enough to understand that water is composed of two hydrogen atoms and an oxygen atom, the term 'water' still referred to H_2O. According to Boyd, this is because there are (and have been) causal mechanisms that bring it about, over time, that what is predicated of the term 'water' will be approximately true of H_2O; our beliefs will converge on truths that would only be true if water was composed of H_2O (for example, certain spectroscopy results will provide evidence for the claim that water is composed of H_2O, and we will trust these results).

The case of moral terms is intended to be analogous to other natural kind terms. Suppose that on Earth we use the term 'right' to commend actions that are particularly praiseworthy, and the predicate 'right' is causally regulated by a certain natural property – maximizing happiness. This would mean that there are causal mechanisms that make what Earthlings predicate of the term 'right' increasingly true of maximizing happiness over time, even though Earthlings may be unaware of this fact. For example, we will increasingly assert rightness of actions like harvesting hitchhikers' organs and giving up our own projects in favour of supporting famine relief in Bangladesh, and these beliefs will be caused in an

appropriate way by the property of maximizing happiness (or by the entities that instantiate maximizing happiness).

In response to Boyd, Horgan and Timmons ask the following question: can we engineer a "twin" Earth where a different property (for argument's sake, consider universalizability) causally regulates the use of 'right'? On such a Moral Twin Earth, there would have to exist causal mechanisms that make what Twin Earthlings predicate of the term 'right' increasingly true of universalizability over time – for example, that actions like harvesting the organs of hitchhikers are wrong.[35] If we can engineer such a scenario, the word 'right' refers to a different property on Earth than on Moral Twin Earth language, and this would give the wrong sense of the disagreement between speakers from Earth and Moral Twin Earth.

However, the success of this "twinning" argument depends on an affirmative answer to the following question:

> Is it *possible* for causal regulation to allow for different properties to regulate the use of two otherwise intertranslatable moral terms in two languages?

If this were not possible, the Moral Twin Earth argument would be incoherent. So we must convince ourselves that it is possible, and the first step to doing so is to get a better understanding of what it is for a property to causally regulate the use of a term. Boyd helps to explicate causal regulation with the following list of mechanisms that will help to lead us to have increasingly true beliefs regarding what is 'right':

---

[35] For sake of brevity, I'm assuming that all actions that are not right are instead wrong. This isn't necessary for my point, but it will help to make some of the examples more clear.

> Such mechanisms will typically include the existence of procedures which are approximately accurate for recognizing members or instances of *k* (at least for easy cases) and which relevantly govern the use of *t*, the social transmission of certain relevantly approximately true beliefs regarding *k*, formulated as claims about t, a pattern of deference to experts on k with respect to the use of *t*, etc... (195)

The question can then be rephrased this way: can we allow these mechanisms to vary between Earth and Twin Earth without changing what is essential to a successful Twin Earth thought experiment? A Moral Twin Earth "twinning" argument strives to keep as many elements consistent between Earth and Moral Twin Earth as possible, and ideally will only allow one feature to vary between the two worlds:

> [C1] The property that causally regulates the word 'right' should change between Earth and Moral Twin Earth.

Since causal regulation is a feedback loop, rather than a one-way relation from a property to a speaker, it will not be possible for *only* the property responsible for causally regulating the moral term to change from world to world – other facts about the world or the society in question will also need to be altered to accommodate this change.[36] And this could spell trouble for the Moral Twin Earth argument. To see why these additional changes might potentially present a difficulty, consider Putnam's original Twin Earth argument. The intuition that 'water' has a different meaning on Earth than on Twin Earth hinges on the fact that there are no differences between Earth and Twin Earth other than whether H_2O or XYZ is present. If other facts on Twin Earth (such as the speaker's mental states, or some other set

[36] Unless the two causally regulating are causally equivalent (as in the water/H_2O case) – but this doesn't seem to be the kind of difference that the moral naturalist will find helpful.

of facts external to the speaker) were different, we might explain the difference in meaning of the word ‘water’ in their respective languages by appealing to those different facts, rather than the compositional difference between water and H_2O.

However, only some of these facts will be relevant to conclusion we wish to obtain from the Moral Twin Earth scenario. The original Twin Earth experiment showed that the meaning of natural kind terms (at least) was dependent on facts that were not internal to the speaker. The Twin Earth argument could have arrived at this conclusion even if other (apparently irrelevant) facts about the world were changed between Earth and Twin Earth, so long as the speaker wasn’t aware of those facts – perhaps Gatorade was ABC instead of a sugary liquid, for example. In the Moral Twin Earth scenario, we are only trying to preserve the identity of the word in question. So the following conditions must hold in the Moral Twin Earth scenario we have concocted, which I take to guarantee the ‘identity’ of the word between linguistic communities:

> [D1] The syntactical item ‘right’ in each language should be fixed between Earth and Moral Twin Earth.
>
> [D2] The relative role that ‘right’ plays in each language should be fixed between Earth and Moral Twin Earth.
>
> [D3] The prescriptivity of sentences featuring ‘right’ in each language, as well as other formal features of the word ‘right’, should be agreed upon by speakers on both Earth and Moral Twin Earth.

Condition [D1] is not actually so necessary for the Moral Twin Earth argument; we can imagine the Moral Twin Earth language having a syntactical item like ‘blargh’ in place of

'right', but still having two distinct properties regulating the usage of 'right'. But setting this condition will help to remove some of the pragmatic factors that might distort our intuitions about the thought experiment. Conditions [D2] and [D3] are more central to the argument. Regarding condition [D2], it seems important to fix the grammatical features of the language that are relevant to the use of moral expressions; for example that sentences expressing moral claims can be negated in two different places. Condition [D3] is the most important – if the word 'right' does not have prescriptive force on Moral Twin Earth, we would be tempted not to think that the words were eligible for translation to begin with. If we keep these features constant across Earth and Moral Twin Earth, we preserve the original inclination to translate their lithographically identical word 'right' between the two communities.

I claim that we can, in fact, vary the property that causally regulates the use of the term 'right' without altering [D1], [D2] or [D3], and thereby show that the Moral Twin Earth argument undermines any form of moral naturalism grounded in causal regulation semantics. Recall that we have to allow the following sorts of facts about the linguistic communities to vary for different properties *k*:

- procedures for recognizing instances of *k*,
- the transmission of true beliefs regarding *k* (formulated as being about *t*),
- a pattern of deference to experts on *k*

Obviously there are some minimal restrictions on the range of properties *k* which can regulate the use of the term t. It won't do for rightness to be causally regulated by some property that cannot be coherently predicated of actions (like redness). And it's not even clear that all properties that properly apply to actions could causally regulate 'right'; there

might be some properties that can't reasonably be conceived of as prescriptive. Perhaps the property of being performed while standing on one leg, or the property of being 35km from a movie theatre, cannot be prescriptive. But all that the Moral Twin Earth argument requires is a single pair of potential referent properties for 'right', and the pair of *maximizing happiness* and *universalizable* at least appear to do the trick. Let us call the culture whose use of 'right' is regulated by maximizing happiness the *Millians*, and the culture whose use of 'right' is regulated by universalizability the *Kantians*. How will a world inhabited by Millians differ from a world inhabited by Kantians, and more importantly, will it differ in a way that undermines [D1], [D2], or [D3]?

As stipulated, there will be different procedures for recognizing maximizing happiness and universalizablity. Millians who are used to identifying examples of 'right' will be able to distinguish happy psychological states from unhappy psychological states, whereas Kantians will be exceptionally competent game theorists, used to puzzling out the possible ways in which having everyone perform a given action might lead to the deadlock or collapse of an institution of some kind. Because of the normative force of the terms and claims involved, it likely follows that Millians and Kantians will be motivated to perform different actions, and have different institutions in their societies. Millians will occasionally endorse lying, for example, while Kantians will tend to strongly condemn lying, and both communities will structure their social institutions accordingly.

But none of this impinges on [D1], [D2] or [D3]. In spite of these substantial differences, we can still imagine the two communities as possessing in their languages the lithographically identical term 'right', using this word in a prescriptive way, and having the meaning of the other words in their language remain fixed.

To give this claim further plausibility, we can examine the sample explanation that Horgan and Timmons provide for how this difference in causal regulation might come about:

> The differences in causal regulation, we may suppose, are due at least in part to species-wide differences between psychological temperament that distinguish Twin Earthlings from Earthlings. (For instance, perhaps Twin Earthlings tend to experience the sentiment of guilt more readily and more intensely, and tend to experience sympathy less readily and intensively, than do Earthlings.) (245)

If we vary the psychological temperament of the populations involved, need anything about the two communities change with respect to [D1], [D2] or [D3]? It's hard to see how any of these constraints could change between Earth and Moral Twin Earth. The two communities may vary in many of the same ways we've already listed (in fact, the difference in temperament could plausibly lead to the Millian and Kantian communities discussed earlier,) but moral language in both communities will still possess all of the formal features demanded by [D1]-[D3].

3.3 Moral Twin Earth and Bare Descriptivist Semantics

In the last section I argued that the Moral Twin Earth argument shows that Boyd's externalist account of the reference of moral terms is implausible, and it is tempting to think that this conclusion extends to all externalist accounts of **R**. But we should return to the question of whether some internalist account of **R** might avoid both prongs of the Moral Twin Earth argument. Recall the kind of metasemantic theory that was tacitly assumed by Hare's internalist picture of moral language:

> The referent property *p* of a moral term *t* is the property that satisfies a description encapsulating *all* the beliefs about the term among the people who speak the language to which *t* belongs.[37]

If we recall Hare's missionaries and cannibals example, a cannibal community will have (at least) three sets of beliefs about 'right'[38]:

> (C1) *Immediate beliefs about the referent property of 'right':* in this case, that the referent of 'right' is the property of being most productive of scalps
>
> (C2) *Beliefs about what is right (or indirectly, the extension of 'right')*: in this case, for example, that it is right to pursue war as often as possible, in order to increase the number of possible opportunities to take scalps.
>
> (C3) *Beliefs about the formal features of 'right':* for example, that 'right' is a prescriptive term, used to commend actions.[39]

However, we can imagine members of a community of missionaries having different beliefs about the referent of 'right', thereby referring to a different property:

> (M1) *Immediate beliefs about the referent property of 'right':* in this case, that the referent of 'right' is the property of being consistent with the Ten Commandments.

---

[37] Jackson (1998) extends this point by claiming that it should involve the mature folk platitudes of the society, where 'mature' is arrived at through some converging process. And it would be tricky to understand how a whole linguistic community could have beliefs about a certain term; clearly the beliefs will vary from person to person and have to be "averaged" in some way. But I take it that this is not a serious problem.

[38] This division of beliefs into three categories is not necessarily exhaustive. A further set would be beliefs about rightness that are substantive and not formal, but which do not involve the extension of the term.

[39] By 'formal features' I mean features about the expressions or properties that are independent of whatever substantive theory of rightness you might hold.

(M2) *Beliefs about what is right (or indirectly, the extension of 'right')*:: in this case, for example, that it is right to obey one's parents.

(M3) *Beliefs about the formal features of 'right':* for example, that 'right' is a prescriptive term, used to commend actions.

Note that (C1) is different from (M1), (C2) is different from (M2), but (C3) and (M3) are constant between the two communities. Because the differences between the two communities are about the substantive theory of rightness, rather than the formal features of the term, it is clear that these differences in beliefs are consistent with conditions [D1]-[D3], which ensure that the Moral Twin Earth thought experiment is coherent:

[D1] The syntactical item 'right' in each language should be fixed between Earth and Moral Twin Earth.

[D2] The relative role that 'right' plays in each language should be fixed between Earth and Moral Twin Earth.

[D3] The prescriptivity of sentences featuring 'right' in each language, as well as other formal features of the word 'right', should be agreed upon by speakers on both Earth and Moral Twin Earth.

Therefore, if we accept the conclusions of the Moral Twin Earth argument, it appears to follow that the internalist picture of language assumed by Hare cannot make sense of our substantive moral disagreements. Moreover, it appears that any internalist theory of reference that allows the reference-fixing beliefs to vary between communities will be undermined by the Moral Twin Earth argument.

In order to solve the Moral Twin Earth problem, we must show that the two communities use 'right' to refer to the same property, and so appeal to only those reference-fixing beliefs that are invariant between communities, given the constraints [D1]-[D3] of the Moral Twin Earth experiment. In particular, we should pay attention to [D3], which tells us that the moral term possesses the same formal features in both communities: for example, that it is prescriptive, used to commend, is concerned with certain kinds of entities that pertain to moral agents (like actions, or the agents themselves) and so forth. This gives the following revised theory of reference:

> The referent property *p* of a moral term *t* is the property that satisfies a description encapsulating *only the beliefs about t that cannot change from world to world,* among the people who speak the language to which *t* belongs.[40]

If [D3] is correct, it seems to follow that the agent is committed to beliefs of the form (C3)/(M3). This isn't an obvious entailment, for it is conceivable that a person might use a term to commend someone without believing of the term that its referent property is prescriptive.[41] However, it seems reasonable to grant this claim. If all this is true, the beliefs of the form (C3)/(M3) will not change from community to community – so the property that satisfies these beliefs will also not vary from community to community. However, there are two reasons to think that this is not an adequate solution to the Moral Twin Earth problem.

---

[40] Jackson (1998) extends this point by claiming that it should involve the mature folk platitudes of the society, where 'mature' is arrived at through some converging process. And it would be tricky to understand how a whole linguistic community could have beliefs about a certain term; clearly the beliefs will vary from person to person and have to be "averaged" in some way. But I take it that this is not a serious problem.

[41] To put the point a different way, perhaps you be an externalist about attitudes (such as commending) as well as beliefs, so that you are not necessarily aware of the attitudes you express in using a term.

Firstly, it seems plausible that we could weaken [D3] even further. There are conceivable examples where competent users of moral terms do not agree upon all of the purely "formal" features of such terms. Consider the following two examples:

a. A community of praisephobic Kantians who strenuously avoid using moral terms to praise either their own or other people's behaviour.
b. A community of amoralists who fail to recognize that moral terms are prescriptive.

Both of these examples might be contested, but if they are plausible that the set of beliefs about the term 'right' that cannot change from community to community is much smaller than originally posited, or is perhaps entirely empty.

Secondly (and more importantly,) it is unclear that purely formal features are enough to fix a unique referent property for a moral term. For example, consider the prescriptive function of moral properties – that an action possessing the property of rightness, for example, gives a person a necessary reason to perform that action. It seems plausible that both maximizing happiness and universalizability are prescriptive in that sense. If not, we need a further explanation of why only one of maximizing happiness or universalizability are prescriptive, and it is unclear how this explanation could be either an analytic or synthetic fact.[42] We should conclude that what I have termed "Bare Descriptivist Semantics" is also undermined by the Moral Twin Earth argument.

[42] Interestingly, I have not presented any reasons that would guarantee that this burden cannot be shouldered. So far we have focused on analytic and synthetic *identities*, but the *predicative* question "Is maximizing happiness prescriptive?" should be explained very differently from questions about an identity between maximizing happiness and prescriptivity. For example, when is it analytically true that we can predicate a property of some entity or property? The Open Question Argument – especially the expanded Open Question Argument, is poorly suited for this task. And what would it be for such identities to be synthetically true?

3.4 The Moral Twin Earth Argument and Analytic Moral Naturalism

Before assessing the Moral Twin Earth argument in the next chapter, it is interesting to note that the Moral Twin Earth argument has important implications for all forms of moral naturalism, including *analytic* moral naturalism. The Moral Twin Earth argument depends on the idea that the relation that fixes the reference of moral expressions (we have been calling this relation **R**) can pick out different referents in different communities. Because synthetic moral naturalism justifies moral-natural identities partly by appealing to contingent facts about the world, the referent of a moral expression could change from linguistic community to linguistic community. We might also ask: is such variation possible according to analytic moral naturalism, where only facts about meaning are responsible for determining the referents of expressions?

The answer to this question hinges on the status that we accord to analytic facts. If analytic truths are facts about the meaning of expressions in their respective languages, then their truth or falsity might depend on the contingent norms governing different languages. If so, it is perfectly reasonable to imagine one culture using 'right' to refer to maximizing happiness, and another culture using 'right' to refer to productive of the most scalps – where these facts are given by the definition of 'right' in each of their languages (on a simple account, perhaps they are facts that can be discovered by looking in a dictionary.)

However, many people think that analytic truths reflect deeper facts about the natural world. Take knowledge, for example. Epistemologists generally proceed as though the conceptual analysis of knowledge into (JTB+X) is an analytic matter, so that armchair data is sufficient to find the correct analysis of knowledge. The hope of these epistemologists is that this data

is relevant beyond cultural and linguistic boundaries – that they really are uncovering truths about a kind (knowledge) that exists prior to cultural and linguistic conventions. However, data from experimental philosophy has undermined these hopes. Many philosophers now agree that people from different cultures have very different intuitions about the cases that constitute knowledge, such as the Gettier problems.[43] These cases threaten to undermine the position that the concept of knowledge is rooted in anything other than the linguistic conventions of a given community.

Therefore, the prospects of analytic moral naturalism for avoiding the Moral Twin Earth argument hinge on a positive answer to this controversial question: can we give non-trivial analyses of concepts that are not merely rooted in cultural and linguistic data? If not, it seems like analytic moral naturalism is also vulnerable to the Moral Twin Earth argument.

[43] See Nichols, Stich, and Weinberg (2003) for a paper that is representative of experimental philosophy regarding Gettier cases.

Chapter 4: Copp's Indirect Argument against Moral Twin Earth

In this chapter, I characterize and discuss David Copp's objection to the Moral Twin Earth argument. First, I show that his precise objection cannot undermine the Moral Twin Earth argument. Second, I argue that the general form of his objection will also not succeed.

4.1 Copp's Indirect Argument

Copp (2000) is responsible for a widely-discussed reply to the Moral Twin Earth argument, and his respondents are best understood as attributing two distinct arguments to him.

First, he advances a referential account **R** that aims to avoid both horns of the Moral Twin Earth argument. Boyd (1988), Brink (1989) and Jackson (1998) present well-known constructive attempts at such an account of **R**. But I think that the considerations presented by Horgan and Timmons (2000) show that Copp's positive account of **R** is not satisfactorily determinate, and I will discuss it no further.

However, Copp's 'second' argument takes a different approach. He gives an *indirect* argument that shows that, regardless of the details of the referential account we have in mind, speakers on Earth and Moral Twin Earth must be referring to the same property. The indirect argument proceeds in this way:

1) The referential intentions of speakers using 'right' at Earth and Moral Twin Earth are identical.
2) The state of the world at Earth and Moral Twin Earth is identical (in the sense that the same moral properties are instantiated in each linguistic community, and the worlds are "the same in all morally relevant aspects" (Copp, 2000)).

3) The reference of 'right' is fixed strictly by the referential intentions of speakers using 'right' and the state of the world.
4) Therefore, people on Earth and Moral Twin Earth refer to the same property with their term 'right'.

Horgan and Timmons will respond, as before, by noting that the underlying account of reference **R** will fail to pick out a determinate referent. This may be so. But Copp's response to this worry is bolstered by a sort of "innocence by association" move. Many other properties that we manage to ostensively refer to in the world, such as health or even (as Copp points out) everyday samples of water, are vague in some sense. This is a problem with characterizing ostensive definitions generally, yet we don't want to close the door on such definitions altogether. After all, description cannot be the only way in which our words hook up to the world – at the risk of regressing, some of the terms that figure into descriptions must have gained their reference through some means other than description, and ostension (or some similar mechanism) is often thought to be the clearest alternative.

In reply, I suspect that Horgan and Timmons have grounds to point out that the ostensive definitions of moral terms involve the *wrong sort* of vagueness. But regardless of the success of this reply, I will question Copp's indirect argument on entirely different grounds. I will argue that given certain plausible assumptions about the nature of referential intentions, there are always Moral Twin Earth scenarios where speakers have divergent referential intentions in using moral terms. But first, in the remainder of this section, I will address the particularities of Copp's argument, which heavily relies upon Putnam's account of the metasemantics of kind terms.

Putnam's Metasemantics for Kind Terms

As I suggested, I take the most favourable reading of Copp's argument to be indirect. Copp decomposes an ostensive definition in a roughly Putnamian way, suggesting that the features relevant to fixing reference are *ostensive referential intentions* and *presuppositions*.[44] If these features do not vary from Earth to Moral Twin Earth, moral terms must pick out the same referent in both linguistic communities.

Putnam thinks that our intentions regarding definitions of natural kinds (and many other common nouns) have an *ostensive character*. For example, in using the word 'water', we point at a sample of H_2O, and intend to refer to the property that makes other samples relevantly similar to that sample of H_2O, namely standing in the *same-liquid* relation to the sample.[45] This ostensive definition has an indexical character, so that in a hypothetical world where H_2O is switched with XYZ, it will turn out that 'water' refers to XYZ for speakers at that world.

Of course, a given sample (or set of samples) can instantiate many different properties; Wayne Gretzky is both a human being and a 10-time Art Ross Trophy winner. We need a mechanism to winnow the field of potential referents to only one property. Putnam suggests that the relevant relation of similarity between the samples depends on the *interests* of the speaker. For example, we use the term 'milk' not with the interest of picking out the essential

---

[44] As far as I can tell, Putnam never explicitly claims that this is an exhaustive list of the features relevant to fixing reference, but Copp needs this claim for his indirect argument to succeed.

[45] Note that in this case, the sortal *liquid* ensures that the ostensive definition does not apply to ice or water vapour, even though they share the same molecular constitution as liquid water. The sortal does not always contribute so directly to the content of the definition; we can imagine relations as vague as *same-thing* or *same-property*.

molecular property that unifies the relevant samples, but instead with the interest of picking out the functional property that unifies them, namely being a liquid produced by mammals to nourish their young.[46] But sometimes the interests of speakers must be complemented with facts about the world to fully explicate the relation of similarity. For example, our interest in using the term 'water' is to pick out the essential property that explains the macroscopic behaviour of the substance, but it is a further *a posteriori* fact that the relevant essential, explanatory property is being H_2O.

Putnam makes some further remarks about the *presuppositions* that such ostensive definitions have. He writes:

> My "ostensive definition" of water has the following empirical presupposition: that the body of liquid I am pointing to bears a certain sameness relation (say, x is the same liquid as y, or x is the same_L as y) to most of the stuff I and other speakers in my community have on other occasions called "water." (142)

These brief remarks about presuppositions are not fully fleshed out, but Putnam appears to be thinking of a *reference-borrowing* use of the word "water", and it is unclear the sense in which one can define a term that is already in common circulation. If the presupposition is not met, Putnam suggests that we would intend to withdraw our definition.

To sum up, in dividing up referential intentions of speakers in a Putnamian way, Copp must show that speakers on Earth and Moral Twin Earth have the following features in common:

---

[46] This example is from Copp (2000).

(A) *Ostensive intentions*: the intention to use the term 'right' to refer to things that bear the indexed *same-X* relation to the sample they are pointing at. This relevant relation between the samples, and therefore the intention, is fixed by three features:

   a. The sortal X in the same-X relation.

   b. The interests of the speaker.

   c. The properties actually instantiated in the world.

(B) *Presuppositions*: the intention to refer to the stuff that their linguistic community have been calling 'right'.

4.2 Responding to Copp's Indirect Argument

I will now argue that we can always construct a Moral Twin Earth scenario in which speakers on Moral Twin Earth have different referential intentions than those on Earth.

The Moral Case: Ostensive Intentions

Regarding the *ostensive intentions* in our use of moral terms, Copp writes the following:

> In using moral terms, we intend to refer to whatever bears the "moral-same-kind relation" to the examples in the world that we point to in explaining what such terms mean. (116)

For example, when defining 'right', people on Earth point at examples of actions that maximize happiness, like harvesting hitchhikers' organs to save multiple cancer patients and giving up luxuries to support famine relief in Bangladesh, and they intend for 'right' to refer to actions similar (in some sense) to those actions. On Moral Twin Earth, people define their orthographically identical word 'right' by pointing at examples of actions that are demanded

by their religious texts, such as observing the Sabbath and performing penance for wrongs, and intend 'right' to refer to actions similar (in some sense) to *those* actions. However, there are two reasons to doubt that the two communities have identical ostensive intentions.

First, we can tweak the above example slightly so that the ostensive definition of Moral Twin Earthers uses samples that *only* instantiate the property of being demanded by a religious text, even though we would want to grant that they are talking about rightness as well. Suppose, for example, that Moral Twin Earth religious texts praised self-immolation before a statue of a deity. Perhaps that is the *only* action that instantiates rightness, on their view. This action, surely, does not fall under the extension of maximizing happiness. Given such a difference in extension between the two expressions, it is hard to see how the ostensive intentions of people on Earth and Moral Twin Earth could ultimately pick out the same property.

Second, even if we grant that speakers from both communities choose similar samples for their definitions, we should doubt whether they have the same sortal *moral-same-kind* in mind, or whether they have the same interests. We lack any positive account of what an interest is, or how to tell if two sets of interests are identical. This is particularly troublesome because we expect that our interests in using a term should be independent from our beliefs about its referent. For example, even if the ancient Greeks believed that gold is a compound, we can understand their word 'gold' as referring to the same property as our word 'gold', even though modern chemistry tells us that gold is an element. We have vastly different beliefs about gold, but the interests of speakers in both eras are roughly clear: we both intend to pick out a certain property that explains the macroscopic properties of a substance with which we interact. However, in the moral case it seems harder to disentangle our interests

from our beliefs about rightness. For example, if our interest in using 'right' is to pick out the property that "makes objects of assessment interpersonally justifiable" (Brink, 2001), this seems like a substantial commitment to a certain claim about rightness – namely that it makes objects of assessment interpersonally justifiable.

Both of these points can be made more compelling by appealing to natural kind cases where our intuitions are pulled away from common reference. Unlike the case of gold, the definition of an acid presents a fairly clear case of reference discontinuity.[47] On the Brønsted-Lowry theory of acidity, an acid is a proton donator, but on the Lewis theory, an acid is an electron acceptor. My intuition on this case is that the interests of both concept users are roughly the same: to pick out a property that explains why substances react to other substances in certain ways. We have the same sortal, similar interests, and overwhelming overlap in samples (there are only a few Lewis acids that are not Brønsted-Lowry acids,) so Putnam's theory predicts that speakers using both terms should refer to the same property. However, I have the intuition that chemists would be picking out two distinct properties. If Lewis claimed that a magnesium ion was an acid, but Brønsted claimed that it was not an acid, as predicted by each of their theories, I wouldn't take them to be in genuine disagreement with one another. We might think that the definition of rightness is more like this case than any of the other examples that Copp presents.

The Moral Case: Presuppositions

I have argued that people on Earth and Moral Twin Earth need not share their *ostensive intentions*. However, Copp takes his major contribution to the debate to be his discussion of

[47] This example is due to Stanford and Kitcher (2000).

presuppositions, and it is still possible that the similarity in presuppositions between speakers "cancels out" the difference in ostensive intentions between speakers. Copp considers which presuppositions an ostensive definition of 'wrong' might carry:

> He would be committed to withdrawing his [ostensive definition] if he came to believe that lying is not of the kind, or does not have the property, that he and most speakers in his linguistic community intend to refer to in using "wrong". (129)

Returning to my earlier example, Copp will want to show that people on Moral Twin Earth are in error about the referent of their terms. Perhaps the incorrect use of the word 'right' on Moral Twin Earth is like accidentally pointing at a sample of gin when one really intended to point at a sample of H_2O in defining 'water' – if so, they would intend to withdraw their definition, and perhaps wish to pick out a different property instead.

To see how this will not help the moral naturalist, note that the presuppositions cited by Putnam rely on some semantic, "reference-borrowing" conventions that are already in place in a linguistic community. If the ostensive definition fails to be consistent with the conventions, this is a fact about the world that would lead a speaker to withdraw their ostensive definition. However, these semantic conventions will differ from linguistic community to linguistic community, so they cannot be used to secure a common referent for moral terms between two different linguistic communities. And without these semantic conventions, their presupposition amounts to this:

> And if he is sincere, he would be committed to withdrawing his [ostensive definition] if he came to believe that lying is not of the kind, or does not have the property, *that he intends to refer to in using "wrong".*

But this is just to say that by 'wrong', he intends to refer to what he intends to refer to in using 'wrong'! We are given no further help in understanding when this presupposition would be activated. I think that we should conclude that the considerations presented in Copp (2000) do not undermine the Moral Twin Earth argument.

4.3 Generalizing Copp's Reply: The Metaphysics of Referential Intentions

I have argued that Copp's particular appeal to referential intentions does not solve the Moral Twin Earth problem. He has not shown that the referential intentions of people on Earth are the same as those of speakers on Moral Twin Earth. But Copp has proceeded by decomposing referential intentions in a particular way, and a moral naturalist might proceed by decomposing referential intentions in an alternative way.[48] If such an alternative decomposition were correct, the moral naturalist could argue that speakers in different linguistic communities actually have the same referential intentions.

In this section, I will argue that this is not possible. To do so, I will have to give a partial account of what is constitutive of *intending to refer* to some object or property using a particular term. It seems clear that intending to refer to something is a state that a person can be in – so we should wonder, what facts about a person (and, conjointly, the world) make it the case that a person is in that state? Or at least, what facts must necessarily hold for a person to be in that state?

One might think that some facts about what the speaker knows, or what the speaker attributes to a referent, would be a precondition for intending to refer to some object or property. For

[48] Here is a rough sketch of one alternative: to have a certain referential intention is just to be in some mental state M when using a term *t*. The moral naturalist could then provide some grounds for thinking that speakers on Earth and Moral Twin Earth are both necessarily in state M.

example, if someone didn't know that the arbitrary and unnecessary causing of harm is wrong, we might think that they cannot use 'wrong' to refer to the same property as a competent user of the concept of wrongness. But a well-known example by Donnellan (1969) demonstrates how such descriptive conditions are likely unnecessary for having a certain referential intention. Imagine that someone was looking at two squares on a wall, square A on top of square B, but was fitted with light-bending glasses that inverted the apparent positions of the squares on the wall. If the person assigned the name 'X' to the square at the top of their field of vision (B), they would be unable to give any true description that actually holds of the square they named 'X'.

This example is sometimes taken to rule out all epistemic conditions on referential intentions, but I don't think that Donnellan's example rules out all such conditions. In particular, I think that there is a counterfactual, epistemic precondition for someone to have an intention to use some term *t* to refer to some object or property X:

> If the speaker knew all the facts about an object or property X, would they agree that their term *t* refers to X?

If the subject does not agree that their term refers to X, they should not be understood as referring to X, though this might not be a *sufficient* condition for them to refer to X. This epistemic condition seems to be an important part of the "intention" in referential intentions; if we are completely unaware of referent of our terms under conditions of full information, it is hard to see how the relevant act of referring could be intentional. To make this test completely clear, I would have to give an account of which facts are *about* an object or property X. I will not do this here, but I intend the range of facts to be understood fairly

broadly.[49] For example, in Donnellan's case of the inverted glasses, had the person known that they were wearing inverted glasses which skewed their perceptual access to the objects in question, they would agree that they intended to use *t* to refer to the square that should have been at the bottom of their vision.

Bearing in mind this precondition on having a certain referential intention, we can return to the following question: how can a person on Moral Twin Earth, who takes herself to use 'right' to refer to being commanded by religious texts, actually *intend* to refer to maximizing happiness? We would have to attribute the following referential intention to her and her linguistic community:

> On Moral Twin Earth, by 'right', people intend to refer to maximizing happiness.

Even if people in a wide variety of linguistic communities have this referential intention, I will show that we can always craft a Moral Twin Earth scenario in which speakers do not share that intention. We can begin by observing that if they have this intention, either it is *direct* or it is *mediated* through some content that fits the referent, such as a description or intension.

If the referential intention is *mediated*, it will be mediated through some sort of description or mental content. The speaker intends to refer to the particular entity that satisfies this description or mental content. This content will either be (a) moral or (b) non-moral in character.

---

[49] Depending on the kind of sceptical scenarios envisioned, we might have to extend the range of facts beyond facts purely about the referent to all of the facts about the natural world. However, I do not think that facts about inter-theory identities (like moral-natural identities,) or facts about the reference relation **R**, should be included in this range of facts.

If the relevant mediate content is *moral*, we will be unable to determine in a non-circular way whether the relevant intentions are the same between two speakers. Several philosophers suggest that speakers have such morally-loaded intentions in using moral terms. In the original 2000 paper, Copp attributes to speakers the intention to pick out the property which is of "primary importance morally in deciding which actions to [do]." (131) On the account of van Roojen (2006), we intend to refer to the property that is "used to reason about considerations bearing on well-being" (172) or the property that is "the most general term of moral appraisal." (173) Brink (2001) echoes this idea by suggesting that language users, in using moral terms, must adopt the "moral point of view" (175).

The difficulty is that we can run a further Moral Twin Earth argument on these descriptions, since what counts as "well-being" or "moral" might vary from linguistic community to linguistic community. So the mere fact that two speakers intend to refer to something that satisfies some moral description does not give us grounds to infer that their referential intentions are identical.

On the other hand, suppose that the relevant description or content is *non-moral*. Copp (2007) suggests such a possibility:

> The basic idea would be that "right" is used with the semantic intention of ascribing to an action or a kind of action the property of being required by the code of rules, whatever it is, the currency of which in the society in question would best contribute to the society's ability to meet its needs. (238)

I would grant Copp that there are no morally-loaded descriptions in this intention; the needs of a society are plausibly a function of some non-moral considerations, like the desires of

people in that society. However, this approach will still not succeed. First of all, like Horgan and Timmons (2000), I suspect that the intention that Copp attributes to speakers is not determinate enough to pick out a single referent property. But even if we grant that the description actually is sufficiently determinate to pick out some single property, such as maximizing happiness, it is not reasonable to attribute that intention to speakers in every linguistic community. The following epistemic test shows that this is not the case:

> If the speaker knew all the facts about maximizing happiness, would they agree that their term 'right' refers to maximizing happiness?

My intuition on this question is that a speaker on Moral Twin Earth, fully informed of the relevant facts about maximizing happiness (and perhaps about their religious texts, as well,) would not change their mind about the referent of 'right'. Perhaps they intended to refer to the property of being demanded by religious texts because it satisfied some description, or perhaps they intended to refer to the property of being demanded by religious texts directly, but in either case, they would have no reason to change their mind about the referent of 'right'. Speakers in some linguistic communities may have the intention to refer to maximizing happiness through some description, in spite of their mistaken beliefs about the satisfier of that description, but we can easily imagine Moral Twin Earth scenarios in which speakers do not have this intention.

The remaining possibility is that the referential intention is *direct*, so that the speaker or linguistic community stands in the *intending-to-refer-to* relation to maximizing happiness without a mediating reference-fixer. It is hard to see what such a direct intention would consist of, but unless it is a brute, *sui generis* fact about referential intentions, I think that the

epistemic test must apply here as well. If the Moral Twin Earth speaker actually intends to refer to the property of maximizing happiness, there must be some sense in which they made a mistake by thinking that they intended to refer to the property of being demanded by religious texts. But the epistemic test shows that they could not be making such a mistake; they really do intend to refer to the property of being demanded by religious texts, and therefore have different referential intentions than speakers on Earth.

4.4 Conclusion

In this chapter, I have argued for two conclusions:

(1) Given a Putnamian decomposition of our referential intentions in using moral terms, like that of Copp (2000), we will be able to construct a Moral Twin Earth scenario in which two linguistic communities have different referential intentions regarding moral terms.

(2) Regardless of how we decompose referential intentions, given a plausible claim about referential intentions, we can *always* construct a Moral Twin Earth scenario in which two linguistic communities have different referential intentions regarding moral terms.

We might be tempted to infer a stronger claim from (2): that we can construct a Moral Twin Earth scenario for *any referential account* ***R***. We might reason to this conclusion in the following way:

1) A referential account **R**, is fixed by referential intentions and the state of the world.
2) The state of the world will not vary from community to community.

3) We can always construct a Moral Twin Earth scenario in which two communities have different referential intentions regarding moral terms.
4) For any referential account **R**, we can construct a Moral Twin Earth scenario in which two communities refer to different properties with their moral terms.

But I think this would be to draw the wrong lesson from my argument. My suspicion is that the whole enterprise of dividing a referential account **R** into referential intentions and the state of the world rests on a mistake. The suggestion that there is some set of facts strictly about what the individual intends to refer to, and another set of facts strictly about the world, is probably too stringent; the two sets are not so neatly separable from one another. So even if we agree that we can always construct a Moral Twin Earth scenario in which two linguistic communities have different referential intentions, it isn't clear that we simultaneously can hold the second premise fixed. If there is any meaningful separation to be had along these lines, the "state of the world" might vary from community to community as well. We should conclude that any indirect response to the Moral Twin Earth argument will have to proceed by identifying non-variable features of the world other than the referential intentions of speaker.

Chapter 5: Conclusion

In this thesis, I have examined the prospects for moral naturalism by examining two of the most prominent arguments that critics have levied against it.

First, I argued that the success of the Open Question Argument depends on the form of analytic moral naturalism under examination. Strongly non-reductive forms of moral naturalism can offer an alternative explanation for the feeling of openness attending questions like "Granted that it is x, is it good?" by appealing to the possible underdetermination of moral concepts.

Second, I examined the Moral Twin Earth argument. If correct, the Moral Twin Earth argument successfully undermines any form of moral naturalism where the moral-natural identity is rooted in *a posteriori* facts that can change between linguistic communities. However, it is unclear whether this is insurmountable for the analytic moral naturalist. One might also question some of the intuitive premises of the Moral Twin Earth argument, especially with regards to what can change between Earth and Moral Twin Earth while maintaining the intuition that the communities are talking past one another.

On the whole, moral naturalism has significant drawbacks, but I hope to have shown that these drawbacks are not as overwhelming as some philosophers have claimed.

Bibliography

Audi, R. 2008. “Intuition, Inference and Rational Disagreement in Ethics,” *Ethical Theory and Practice* 11(5): 475-492

Bedke, M. 2008. “Ethical Intuitions: What They Are, What They Are Not, and How They Justify,” *American Philosophical Quarterly* 45(3): 253-270

Bedke, M. Forthcoming. “Against Normative Naturalism,” *Australasian Journal of Philosophy*

Bedke, M. Unpublished draft. “Introducing GAP Expressivism”

Boyd, R. 1988. “How to be a Moral Realist,” in Sayre-McCord, G (ed.), *Essays on Moral Realism*. Ithaca: Cornell University Press

Brink, D. 1989. *Moral Realism and the Foundation of Ethics.* New York: Cambridge University Press

Brink, D. 2001. “Realism, Naturalism, and Moral Semantics,” *Social Philosophy and Policy* 18: 154–176

Copp, D. 2000. “Milk, Honey, and the Good Life on Moral Twin Earth,” *Synthese* 124(1): 113-137

Copp, D. 2007. *Morality in a Natural World.* New York: Cambridge University Press

Darwall, S., Gibbard, A. And Railton, P. 1992. “Towards Fin de Siecle Ethics: Some Trends,” *The Philosophical Review* 101(1): 115-189

Davidson, D. 1963. “Actions, Reasons and Causes,” *Journal of Philosophy* 60: 685–700

Donnellan, K. 1970. “Proper Names and Identifying Descriptions,” *Synthese* 21: 335-358

Feldman, F. 2005. “The Open Question Argument: What it Isn’t, and What it Is,” *Philosophical Issues* 15: 22-43

Geirsson, H. 2005. “Moral Twin Earth and Moral Semantic Realism,” *Erkenntnis* 62: 253-278

Gibbard, A. Unpublished draft. *Thoughts and Oughts.*

Haidt, J. 2001. “The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment,” *Psychological Review* 108: 814-834

Hare, R.M. 1952. *The Language of Morals.* Cambridge: Oxford University Press

Horgan, T. and Timmons, M. 1992. “Troubles on Moral Twin Earth: Moral Queerness Revived,” *Synthese* 92: 221-260

Horgan, T. and Timmons, M. 2000. “Copping Out on Moral Twin Earth,” *Synthese* 124(1): 139-152

Huemer, M. 2005. *Moral Intuitionism.* New York: Palgrave MacMillan

Huemer, M. 2008. “Revisionary Intuitionism,” *Social Philosophy and Policy* 25: 368-392

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

Kripke, S. 1972. *Naming and Necessity*. Oxford: Blackwell

Merli, D. 2002. “Return to Moral Twin Earth,” *Canadian Journal of Philosophy* 32: 207-240

Moore, G.E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press

Putnam, H. 1975. “The Meaning of Meaning,” *Philosophical Papers, Vol. II : Mind, Language, and Reality*. Cambridge: Cambridge University Press

Quine, W.V.O. 1951. "Two Dogmas of Empiricism". *Philosophical Review* 60 (1): 20–43

Rubin, M. 2008. “Sound Intuitions on Moral Twin Earth, *Philosophical Studies* 139 (3): 307–327

Sayre-McCord, G, (ed.). 1988. *Essays on Moral Realism*. Ithaca: Cornell University Press

Stanford, P.K. and Kitcher, P. 2000. “Refining the Causal Theory of Reference for Natural Kind Terms,” *Philosophical Studies* 97: 99-129

Sturgeon, N. 1988. “Moral Explanations,” in Sayre-McCord 1988: 229–255

Van Roojen, M. 2007. "Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument," *Oxford Studies in Metaethics, Volume I* ed. by Russ Shafer-Landau, Oxford: Oxford University Press