University of Massachusetts Amherst

Synthetic Ethical Naturalism

Item Type	dissertation
Authors	Rubin, Michael
DOI	10.7275/qhda-m972
Download date	2025-04-15 18:22:54
Link to Item	https://hdl.handle.net/20.500.14394/38676

SYNTHETIC ETHICAL NATURALISM

A Dissertation Presented

by

MICHAEL RUBIN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2009

Philosophy

© Copyright by Michael Rubin 2009

All Rights Reserved

SYNTHETIC ETHICAL NATURALISM

A Dissertation Presented

by

MICHAEL RUBIN

Approved as to style and content by:

_____
Fred Feldman, Chair

_____
Phillip Bricker, Member

_____
Hilary Kornblith, Member

_____
Christopher Potts, Member

_____
Phillip Bricker, Department Head
Philosophy

DEDICATION

To my family: Linda, Michelle and Neil Rubin

ACKNOWLEDGEMENTS

In writing this dissertation, I have incurred a number of debts, not all of which are financial. My greatest debt is to my dissertation director Fred Feldman. I am also grateful to Hilary Kornblith, Jason Raibley, Lynne Rudder Baker, Jake Bridge, Phil Bricker, Vere Chappell, Sam Cowling, Dan Doviak, Chris Heathwood, Kristen Hine, Justin Klocksiem, Uri Leibowitz, Helen Majewski, Kris McDaniel, Kirk Michaelian, Andrew Platt, Chris Potts, Alex Sarch, Kelly Trogdon, and Brandt Van der Gaast.

Material from chapters one and two appear in "Sound Intuitions on Moral Twin Earth," *Philosophical Studies* (2008) 139: 307-327. I thank Springer for permission to reprint that material here. Chapter five is published as "Is Goodness a Homeostatic Property Cluster?" (2008) *Ethics* 118: 496-528. I thank the University of Chicago Press for permission to reprint that material here. I am grateful to anonymous referees from both journals for helpful comments.

ABSTRACT

SYNTHETIC ETHICAL NATURALISM

FEBRUARY 2009

MICHAEL RUBIN, B.A., BOSTON UNIVERSITY

M.A., NEW YORK UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Fred Feldman

This dissertation is a critique of synthetic ethical naturalism (SEN). SEN is a view in metaethics that comprises three key theses: first, there are moral properties and facts that are independent of the beliefs and attitudes of moral appraisers (moral realism); second, moral properties and facts are identical to (or constituted only by) natural properties and facts (ethical naturalism); and third, sentences used to assert identity or constitution relations between moral and natural properties are expressions of synthetic, *a posteriori* necessities. The last of these theses, which distinguishes SEN from other forms of ethical naturalism, is supported by a fourth: the semantic contents of the central moral predicates such as 'morally right' and 'morally good' are fixed in part by features external to the minds of speakers (moral semantic externalism).

Chapter 1 introduces SEN and discusses the most common motivations for accepting it. The next three chapters discuss the influential "Moral Twin Earth" argument against moral semantic externalism. In Chapter 2, I defend this argument from the charge that the thought experiment upon which it depends is defective. In Chapters 3 and 4, I consider two attempts to amend SEN so as to render it immune to the Moral

Twin Earth argument. I show that each of these proposed amendments amounts to an abandonment of SEN.

Chapter Five explores Richard Boyd's proposal that moral goodness is a "homeostatic property cluster." If true, Boyd's hypothesis could be used to support several metaphysical, epistemological, and semantic claims made on behalf of SEN. I advance three arguments against this account of moral goodness.

In the sixth chapter, I argue that moral facts are not needed in the best *a posteriori* explanations of our moral beliefs and moral sensibility. Because of this, those who accept a metaphysical naturalism ought to deny the existence of such facts or else accept skepticism about moral knowledge. In Chapter 7, I consider a counterargument on behalf of SEN to the effect that moral facts are needed in order to explain the predictive success of our best moral theories. I show that this argument fails.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

ABSTRACT vi

CHAPTER

1. MORAL REALISM, ETHICAL NATURALISM, AND THE NECESSARY *A POSTERIORI* 1

1.1. Introduction. 1
1.2. Moral Realism. 4

1.2.1. Moral Realism and Moral Constructivism. 4
1.2.2. The presumptive case for moral realism. 6

1.3. Naturalism: Metaphysical and Ethical. 9

1.3.1. Ethical naturalism. 9
1.3.2. Metaphysical naturalism. 11

1.4. Analytic Ethical Naturalism. 12

1.4.1. The general strategy of analytic ethical naturalism. 12
1.4.2. Semantic assumptions of analytic ethical naturalism. 13
1.4.3. First-order normative ethics as conceptual analysis. 16

1.5. The Rejection of Realistic Analytic Ethical Naturalism. 17

1.5.1. The rejection of analyticity. 17
1.5.2. Doubts about descriptivism. 18
1.5.3. The open question argument. 19
1.5.4. Chauvinistic conceptual relativism. 21
1.5.5. The argument from normativity. 23
1.5.6. Analyticity and stance-independence. 25

1.6. Synthetic Ethical Naturalism. 26

1.6.1. The necessary *a posteriori*. 26
1.6.2. The semantic and metaphysical underpinnings of the necessary *a posteriori*. 28
1.6.3. Causal theories of reference. 30
1.6.4. The epistemological commitments of SEN. 34

1.7. How SEN answers the objections to AEN.....39

2. MORAL TWIN EARTH VERSUS EXTERNALIST MORAL SEMANTICS.....43

2.1. Introduction.....43
2.2. Putnam's Twin Earth. 44
2.3. Moral Twin Earth.....47
2.4. The Attack on the Moral Twin Earth Thought Experiment.....51

2.4.1. Introduction.....51
2.4.2. A preliminary objection. 52
2.4.3. First objection: competing theories of a kind. 55
2.4.4. First reply. 56
2.4.5. Second reply.....57

2.5. Isolating Moral Properties.....60

2.5.1. Second objection. 60
2.5.2. Reply to (i). 62
2.5.3. Reply to (ii). 63
2.5.4. Reply to (iii). 65

2.6. Functional and Non-Functional Kinds. 68

2.6.1. Third objection. 68
2.6.2. Reply. 69

2.7. Conclusion. 70

3. MORAL TWIN EARTH AND HIGHER-LEVEL PROPERTIES 72

3.1. Introduction. 72
3.2. A Prelimary Sketch of the Higher-Level Properties Reply to Moral Twin Earth.....73
3.3. Functionalism about Mental Properties. 75
3.4. The Higher-Level Properties Reply to MTE. 78
3.5. Troubles for the Higher-Level Property Reply: Agent's-Group Moral Relativism. 81
3.6. "Merely Possible" Relativism. 84
3.7. Is Agent-Relativism Compatible With Moral Realism? 86
3.8. Conclusion. 93

4. BRINK'S MORAL SEMANTICS 94

4.1. Introduction. 94
4.2. Brink's Moral Semantics. 95

4.2.1. Brink's Moral Semantics: an initial formulation. 95
4.2.2. A revised formulation of Brink's Moral Semantics. 98

4.3. Interpersonal Justification. 101
4.4. First Horn: Internal Reasons. 104

4.4.1. Internal reasons and the failure to converge. 104
4.4.2. Smith's absolutist conception of internal reasons. 105
4.4.3. Internal reasons and the stance-dependence of normative facts. 107

4.5. Second Horn: External Reasons. 109

4.5.1. BMS, IJ, and external reasons. 109
4.5.2. Naturalism and external reasons. 111
4.5.3. Reductive accounts of external reasons. 118

4.6. Conclusion. 119

5. IS GOODNESS A HOMEOSTATIC PROPERTY CLUSTER? 121

5.1. Introduction. 121
5.2. Boyd's Homeostatic Property Cluster Kinds 123

5.2.1. Homeostatic Property Clusters. 123
5.2.2. Homeostatic Property Cluster Kinds. 125
5.2.3. Two Examples of HPC Kinds. 128

5.3. Homeostatic Consequentialism: THE MORAL GOOD as an HPC Kind. 131
5.4. The Case against Homeostatic Consequentialism. 135

5.4.1. Isolated Goods. 135
5.4.2. An alternative formulation. 137
5.4.3. Two structural disanalogies. 140
5.4.4. An alternative cluster of properties. 142

5.5. Inductive Inference and THE GOOD. 145

5.5.1. Outline of the Argument. 145
5.5.2. Biological kinds versus moral kinds, part I. 146
5.5.3. Biological kinds versus moral kinds, part II. 149
5.5.4. Social kinds. 152

5.6. An Anticipated Rebuttal. 157
5.7. Conclusion. 161

6. THE EXPLANATORY IMPOTENCE OF MORAL FACTS 162

6.1. Introduction. 162
6.2. The Harman-Sturgeon Exchange. 165

6.2.1. Harman's opening salvo. 165
6.2.2. Sturgeon's reply to Harman. 167
6.2.3. Sturgeon's *tu quoque* reply. 169
6.2.4. Lessons of the Harman and Sturgeon exchange. 171
6.2.5. Moral explanations of non-doxastic phenomena. 173

6.3. Anti-Realist Explanations of Moral Theory. 176

6.3.1. Non-moral explanations. 176
6.3.2. A Darwinian account of moral sensibility. 179
6.3.3. From the Darwinian account to moral anti-realism. 185
6.3.4. Evidence favoring the Darwinian explanation of moral theory.....186

6.4. Return of the Tu Quoque?.....188

6.4.1. Tracking accounts of moral sensibility. 188
6.4.2. Debunking explanations of scientific thought. 191
6.4.3. Railton's evolutionary tu quoque. 192
6.4.4. The social-historical case against scientific realism. 194

6.5. Breaking the Tu Quoque: The Case for Scientific Realism.....199

6.5.1. Overview. 199
6.5.2. The standard case for scientific realism. 199
6.5.3. Approximate truth. 202
6.5.4. Non-epistemic methodological principles and Boyd's version of the ultimate argument. 204
6.5.5. What if the ultimate argument is a failure?.....207

6.6. Conclusion. 209

7. THE PROSPECTS FOR AN ULTIMATE ARGUMENT FOR MORAL REALISM.....211

7.1. Introduction. 211
7.2. Moral Theories and Empirical Predictions. 212

7.2.1. On the instrumental reliability of moral theories. 212
7.2.2. A bad argument against the instrumental reliability of moral theories. 213
7.2.3. Three examples of prediction by moral theory. 214

7.3. Example A: Predictions Grounded in Two Deontic Moral Principles. 219

7.3.1. The implausibility of premise A2. 219
7.3.2. AUh predicts unsuccessfully. 220
7.3.3. Example A restated using a different moral theory. 221
7.3.4. The approximate truth of OU is not needed to explain its predictive success. 225

7.4. Example B: Predictions Based on Moral Character. 226

7.4.1. The independence of the prediction. 226
7.4.2. Trouble from social psychology. 228
7.4.3. A competing non-moral explanation of the predictive success of B1. 231

7.5. Example C: Predictions Grounded in a Causal-Moral Generalization. 236

7.5.1. Does justice really cause social stability? 236
7.5.2. Is C2 non-negotiable condition of adequacy for theories of justice? 237
7.5.3. Justice is not what explains stability. 239

7.6. Moral Explanations and Interesting Generalizations. 240
7.7. Conclusion. 246

APPENDIX: A DEFENSE OF MORAL TWIN EARTH FROM MISCELLANEOUS OBJECTIONS 248

BIBLIOGRAPHY 266

CHAPTER 1

MORAL REALISM, ETHICAL NATURALISM, AND THE NECESSARY *A POSTERIORI*

1.1. Introduction.

Synthetic ethical naturalism (SEN) is a theory in metaethics according to which: (1) there are stance-independent moral properties and facts; (2) these properties and facts are identical to—or otherwise constituted only by—natural properties and facts; and (3) sentences used to assert identity or constitution relations between moral and natural properties and facts express synthetic *a posteriori* necessities. The goal of the present chapter is to explain more clearly what these three theses amount to and to present the central considerations that make SEN a *prima facie* attractive metaethical view. In sketching the commitments of SEN, I will be deferring primarily to the work of three of its early proponents: Richard Boyd, David Brink, and Nicholas Sturgeon.[1] In the subsequent chapters of this dissertation I will argue that we should reject SEN.

It may be worthwhile to take a moment to get clear about the goals of metaethical inquiry. I find that the best way to do this is by contrasting metaethics with what is sometimes called *first-order ethics* or *normative ethics*. It is difficult to give a general

[1] A fourth that deserves mention is Peter Railton. Although others seem to view Railton as holding the same sort of view as the trio just mentioned, his metaethical outlook is different enough to warrant hesitation about lumping him in with this group. Most importantly, Railton—unlike Boyd, Brink, and Sturgeon—expresses ambivalence as to whether moral realism really requires that moral facts be stance-independent (see his 1995 and 1996; for a characterization of stance-independence, see §1.2.1 below). On top of this, he seems favorably disposed towards an ideal observer kind of metaethical theory. (This is most explicit in his 1996.) Because ideal observer theories render moral facts stance-dependent, to the extent that Railton accepts such a view he should not be counted as a moral realist. At any rate, he does not appear to be a realist in the same robust sense in which Boyd, Brink, and Sturgeon are. In spite of all this, Railton's credentials as a metaphysical naturalist are impeccable and I will appeal to his work when explaining the ontological commitments of ethical naturalism.

description of first-order ethical claims without begging any questions against a particular metaethical view.[2] It is best to proceed, then, by way of example. The following sentences express first-order ethical claims: 'Ann is morally obligated to fulfill her promise to Ben'; 'It is good that Carl is happy'; 'Dana is a virtuous person.' While these examples are all expressions of *particular* moral judgments, first-order ethics includes claims of a more general kind, such as 'lying is morally wrong' and 'pleasure is better than pain.' In addition, the subject matter of first-order normative ethics includes general moral theories. Among these are theories in the normative ethics of behavior, which are intended to tell us what makes a morally right action morally right; theories in axiology, which are intended to tell us what makes one life, state of affairs, or possible world non-instrumentally better than another; and theories in virtue ethics, which are intended to tell us which traits of character make an agent virtuous.

It will be useful to have before us some sample first-order ethical theories to refer to. Here are two historically important theories in the normative ethics of behavior:

AUh: Necessarily, for any act-token, x, x is morally right iff x maximizes hedonic utility (i.e., the balance of pleasure over pain).

CI-2: Necessarily, for any act-token, x, x is morally right iff the agent of x, by performing x, treats no person as a mere means.

AUh represents a hedonistic form of act-utilitarianism. CI-2 is a restatement of the second formulation of Immanuel Kant's categorical imperative. It should be noted that both AUh and CI-2 are more than just *theories* about what makes a morally right action right; each expresses a *standard* of morally right action that one may accept or reject.

[2] If I were to set aside concerns about maintaining metaethical neutrality, I would describe first-order ethics in this way: a first-order ethical claim ascribes a normative, evaluative, or moral property, such as *moral wrongness*, *goodness*, or *virtuousness* to an action (or kind of action), state of affairs, or character trait etc. The trouble with this characterization is it precludes a classical non-cognitivist account of moral judgments.

That is to say, each theory doubles as a possible moral code that agents or social groups could adopt to regulate their behavior.[3]

Metaethical theories are theories about first-order ethical claims. Metaethical inquiry includes (but is not limited to) questions concerning the semantic, metaphysical and epistemological commitments of first-order ethical thought and discourse. Among the semantic questions posed by metaethical inquiry are these: Do moral utterances and sentences express truth-apt propositions? If they do, what is the meaning or semantic content of moral predicates? If the primary function of moral utterances is not to express propositions, do they have some other important function? The core metaphysical questions addressed by metaethical inquiry include (i) the question of whether there are moral facts or true moral propositions, (ii) the question of whether moral properties and facts are natural, supernatural, or non-natural, and (iii) the question of whether such facts are "objective" or "stance-independent." Epistemological questions asked by metaethicists include these: Do we have any moral knowledge (or, at any rate, epistemically justified moral beliefs)? If we do, are moral truths discoverable *a priori*, or only *a posteriori*? Are our moral beliefs justified by their being inferable from foundational or epistemically basic propositions, or are they justified in virtue of their mutual coherence with the rest of our beliefs?

[3] I hesitate to say that this will be true of all moral theories. It might be argued that a bare divine command theory in the normative ethics of behavior—a theory according to which an act is morally right just in case it is permitted by God's commands—does not, by itself, constitute a moral code, since, without a further description of God's commands, it cannot be used by agents in any meaningful way to regulate their conduct. It is not important for my purposes that I take a stand on this matter. It will suffice that some theories in the normative ethics of behavior, including AUh and CI-2, double as standards of conduct. My purpose in bringing this double nature of moral theories to the reader's attention is to warn that I will sometimes speak of AUh and CI-2 as theories of right action and at other times speak of them as standards of conduct or moral codes.

1.2. Moral Realism.

1.2.1. Moral Realism and Moral Constructivism.

SEN, as I understand it, is both a form moral cognitivism and a form of moral realism. Moral cognitivists take sentences of the form, 'ϕ is morally right,' to express propositions. Such propositions involve the ascription of a property (*viz*., *moral rightness*[4]) to act-tokens and are truth-evaluable in a straightforward way. By contrast, according to traditional versions of non-cognitivism, the primary function of moral sentences and utterances is to express prescriptions or attitudes, rather than truth-evaluable propositions.[5]

Moral realism is the view that (1) there are moral properties and facts and (2) these properties and facts are "stance-independent."[6],[7] The first clause distinguishes moral realism from moral nihilism. The second clause, which we may call 'the stance-independence clause,' distinguishes moral realism from moral constructivism. The most clear and concise characterization of the stance-independence clause that I know of belongs to Russ Shafer-Landau. By his characterization, stance-independence requires

---

[4] Throughout this dissertation I italicize terms that refer to properties (e.g. *redness*, *roundness* and *rightness*). I do not italicize the names of properties when these names are mentioned and not used; in such cases, those terms refer to linguistic items rather than to properties. Finally, I do not italicize terms referring to property instances or tropes (e.g., the redness of Ann's shirt). In Chapter 5, which deals more directly with natural kinds, I add a convention of using small capital letters for names of kinds (such as GOLD, THE TIGER, and BACHELORS). Although the metaphysics of kinds is controversial, it may help the reader to know that I tend to think of kinds as a distinct sort of entity from their corresponding properties. As I see it, the property of *being a tiger* stands in the same relation to the kind THE TIGER that the property *being Socrates* stands to the man Socrates.

[5] This is a very simplified description of non-cognitivism. Although a non-cognitivist must hold that the *primary* function of moral sentences (e.g. 'ϕ is morally right') is to express an attitude or to issue a prescription, some non-cognitivists allow that moral sentences may have a secondary, descriptive function. (See, for example, Hare 1952: ch. 7). Furthermore, contemporary non-cognitivists have appealed to minimalist theories of truth in order to justify the predication of *truth* to moral sentences (see Blackburn 1998: 75-83).

[6] The term 'stance-independence' was introduced by Russ Shafer-Landau (2003: 15) who credits Ron Milo.

[7] Sturgeon and Boyd, but not Brink, add a third, epistemological component to their statements of moral realism: "…our ordinary methods of arriving at moral judgments provide us with at least some approximate knowledge of moral truths" (Sturgeon, 1986b: 117; cf. Boyd 1988: 182).

that "the moral standards that fix the moral facts are not made true by virtue of their ratification from within any given actual or hypothetical perspective" (2003: 15).[8],[9] Thus, a relativist metaethical view according to which an action's being morally right consists in its being permitted by the moral code that is accepted by the members of the agent's society fails to render moral facts stance-independent, and therefore, should be construed as a form of moral constructivism, rather than a form moral realism. Note that the same holds for ideal observer views of the kind proposed by Roderick Firth (1952). According to a view of this sort, an action's being morally wrong consists in the fact that it violates the moral code that would be endorsed every observer in suitably idealized epistemic conditions. If it should turn out that all ideal observers would endorse the same moral code, then the ideal observer view would vindicate ethical absolutism. Even so, it would not be a genuinely realist metaethical position since the moral standard that fixes the moral facts is made true by the fact that ideal observers would endorse or ratify it.

It is worth mentioning here a third form of moral constructivism according to which the truth of a moral theory T simply consists in the fact that T is the moral theory that we would believe or accept, were our beliefs to achieve a state of reflective equilibrium or maximal coherence.[10],[11] What is noteworthy about this form of constructivism is that it incorporates the same coherentist moral epistemology that is accepted by all SEN proponents (Boyd 1988; Brink 1989: Ch. 5; Sturgeon 2002). For the moral realist, however, coherence reasoning is seen as a procedure for *discovering* the

[8] Italicized in the original.
[9] Boyd, Brink, and Sturgeon all include a stance-independence clause in their own formulations of moral realism (Boyd 1988: 182; Brink 1989: 17; 2001: 154; Sturgeon 1986b: 117).
[10] John Rawls defends a form of constructivism along these lines in his (1980). However, because he is reluctant to ascribe *truth* to substantive moral theories, he might not accept this particular formulation.
[11] For an account of the method of reflective equilibrium in moral theorizing, see Rawls (1971/1999: 40-46) and Daniels (1979).

moral facts. Among other things, the realist takes it to be a logical possibility that the moral theory that we would accept in reflective equilibrium (and, so, are justified in believing) is nevertheless false.[12] The moral constructivist, by contrast, views the procedure of coherence reasoning as by itself settling what the moral facts are;[13] the theory that we would accept in reflective equilibrium is *by that very fact* the true theory.

1.2.2. <u>The presumptive case for moral realism.</u>

It is has been claimed that the conjunction of moral cognitivism and moral realism accords with commonsense ethical thought better than its metaethical rivals do. This is taken as grounds for thinking of cognitivist moral realism as the default metaethical position: the burden of argument is on opponents of cognitivism and moral realism to show that these views fail or that rival metaethical views are superior (Brink 1989: ch. 2). I think this is right. I want to briefly outline the main considerations that support the claim that cognitivism and moral realism best accord with commonsense ethical thought.

At least four considerations serve as *prima facie* evidence favoring moral cognitivism over non-cognitivism.[14] First, the surface grammar of moral sentences (and utterances) is declarative. This suggests that the primary use of moral sentences is to express propositions. Second, speakers of our language often ascribe truth and falsity to moral sentences. By the most natural way of understanding what it is for a sentence to be true, a sentence is true just in case it expresses a proposition that is true. Thus, the

[12] This characteristic of moral realism—that it allows for the possibility that the theory we would accept under ideal epistemic conditions is false—is stressed by Brink (1989: 31-36). Compare this with Putnam's characterization of metaphysical (rather than moral) realism as "radically non-epistemic" in the sense that "the theory that is 'ideal' from the point of view of operational utility, inner beauty and elegance, 'plausibility', simplicity, 'conservatism', etc., *might be false*" (Putnam 1977: 485).

[13] Or, to be more precise, the constructivist views the would-be results of coherence reasoning as settling by itself what the supervenience bases of moral properties are.

[14] In this paragraph, I draw on Brink (1999: 196-199) and Shafer-Landau (2003: 23f).

practice of ascribing truth or falsity to moral sentences suggests that commonsense thought takes such sentences to express propositions, as cognitivists claim. Third, moral sentences can appear in unasserted contexts; for instance, they can be embedded in the antecedent of a conditional statement. In such contexts it is not plausible to claim that the embedded moral sentence functions to express a prescription or attitude: the person who sincerely utters 'if abortion is wrong, then emergency contraception is wrong as well,' for example, does not thereby express an attitude of disapproval towards acts of abortion. Standard non-cognitivist views according to which moral utterances express the speaker's attitudes have trouble accounting for this feature of moral discourse (Geach 1960; 1965). Cognitivism, by contrast, has no trouble accounting for it. Fourth, moral sentences appear as premises in seemingly valid deductive arguments. If moral sentences express propositions, then there is no trouble in seeing how to accommodate the validity of such arguments. If moral sentences express attitudes, however, then things are more difficult; unless we are willing reject the validity of moral arguments as a *mere* appearance, a logic of attitudes needs to be constructed (Geach *ibid*.). Since attempts to construct a logic of attitudes have proven to be controversial, this consideration *prima facie* favors moral cognitivism, which can account for logical relations between moral statements using the widely accepted resources of propositional logic.[15]

Commonsense moral thought, then, seems to favor moral cognitivism. What about moral realism? It should be uncontroversial, assuming cognitivism, that the default commonsense view is that there are moral facts. Since virtually everyone makes moral claims, it stands to reason that it is part of commonsense moral thinking that some of

[15] For a look at attempts to articulate a logic of attitudes, see Blackburn (1984; 1998) and Gibbard (1990; 2003).

those claims are true. Of course, it might be that speakers use moral discourse in order to describe a useful fiction, as some moral fictionalists have claimed. But even proponents of such a view admit that moral fictionalism does not capture moral discourse as laypersons actually use it, but rather describes a way in which we might continue to use moral discourse after we have come to accept (on the basis of philosophical argument) that all affirmative first-order moral claims are literally false (Joyce 2001: 185f; Nolan et al. 2005: 309).

If moral realism is to be credited as the default metaethical position of commonsense, it is not enough that moral discourse contains an implicit commitment to moral facts; it must also include a commitment to the stance-independence of such facts. Here, I think matters are more difficult. My own pre-theoretical intuitions support moral realism on this score: when I am struck by a moral intuition, I normally experience it as an appearance of a fact that obtains independent of what any appraiser (real or imagined) may happen to think. Others, however, report a different experience of moral value and obligation. Gilbert Harman writes,

> I have always been a moral relativist. As far back as I can remember thinking about it, it has seemed obvious to me that the dictates of morality arise from some sort of convention or understanding among people, that different people arrive at different understandings, and that there are no basic moral demands that apply to everyone. For many years, this seemed so obvious to me that I assumed it was everyone's instinctive view, at least everyone who gave the matter any thought in "this day and age" (1985: 27).

Harman and others (e.g. Nichols 2004: 169f; Stich and Weinberg 2001: 641) also note that a significant number of college undergraduates profess their acceptance of moral relativism. Because standard versions of moral relativism take the moral rightness of actions to depend upon the attitudes of people belonging to certain social groups, such

views violate moral realism's stance-independence clause. Consequently, unless we have reason to think that Harman and relativist college students are atypical, we might lack justification for taking moral realism to be the default metaethical position of commonsense.

Fortunately for those who assert realism to be the metaethics of commonsense, there are studies that purport to show that children regard the moral wrongness of certain actions as independent of human conventions and responses (see Nichols 2004: 167-177). These findings suggest that laypersons espousing moral relativism do so as a matter of their moral education—rather than on the basis of unvarnished moral appearances. While these studies and their conclusions could be contested, I am inclined to grant that commonsense moral theory includes a claim to the stance-independence of moral facts. In any case, since my goal in this dissertation is to argue that SEN's brand of moral realism should be rejected, I see no harm in granting that moral realism enjoys a default status in metaethics while the burden of argument rests on its opponents.

1.3. Naturalism: Metaphysical and Ethical.

1.3.1. Ethical naturalism.

It is easy enough to say what ethical naturalism is: it is the view that moral, normative, and evaluative properties and facts are identical with, or else constituted only by, natural properties and facts (cf. Brink 1989: 22, 176ff; Sturgeon 2006b: 92).[16] What is not so

[16] The constitution relation is discussed in Brink (1989: 157-160) and Sturgeon (1986a: 75). Constitution has two jobs to perform for the ethical naturalist. First, it explains the supervenience of the moral on the natural in a way that does not require moral facts to be anything "over and above" natural facts. Thus, it avoids the need for a kind of epiphenomenalist account of the supervenience relation between the moral and the natural. Second, it permits moral properties to be "multiply realizable." The claim that moral properties are multiply realizable is thought to absolve the ethical naturalist of the need to promise that

easy is to say what a natural property or fact is. Some metaethicists have counted as many as seven distinct ways the natural/non-natural distinction for properties has been drawn (Copp 2003; Ridge 2008). This is not the space to canvass every proposal that has been offered. To my mind, the most promising conceptions of *natural propertyhood* invoke an epistemological criterion. Drawing on David Copp's (2003), the account that I favor is roughly this:

> NP: a property, P, is natural just in case a synthetic proposition to the effect that an individual instantiates P can be known only by way of empirical investigation, if it can be known at all.[17],[18]

By NP, a property such as *maximizing the balance of pleasure over pain* counts as a natural property since the question of whether an act-token instantiates this property can be settled, if at all, only by empirical means such as observation, induction and perhaps inference to best explanation. Although I am not in possession of a precise account of

---

moral properties are reducible to natural properties. For brevity, I will focus on identity claims offered by naturalists, leaving aside constitution claims.

[17] Here I follow the provisional formulation of *natural propertyhood* that appears in Copp's (2003: 185f). I am somewhat hesitant to sign on for the refinements that Copp proposes to this formulation. My hope is that NP will serve us well enough.

[18] Compare NP with other accounts offered by metaethicists: "According to [Naturalistic Ethics], Ethics is an empirical or positive science: its conclusions could be all established by means of empirical observation and induction" (Moore 1903); "What I find plausible (even if not a conceptual truth) is that ethical facts could not be natural if they could not be investigated empirically" (Sturgeon 2003: 543n24); "Natural facts and properties are presumably something like those facts and properties as picked out and studied by the natural and social sciences (broadly conceived)…" (Brink 1989: 22); "The natural is whatever is the object of study by the natural sciences. […] [A] science is a natural science just in case its fundamental principles are discoverable a posteriori, through reliance primarily on empirical evidence." (Shafer-Landau 2006: 212, 213); "The vague, pre-theoretic idea that the philosophical naturalist tries to articulate and defend is that everything – including any particulars, events, facts, properties, and so on – is a part of the natural, physical world that science investigates" (Timmons 1999: 12).

I should address an objection to criteria of *natural propertyhood* such as NP. While he accepts that a property's being amenable to empirical investigation is *necessary* for its counting as natural, Sturgeon expresses doubt as to whether this is *sufficient* on the grounds that there could in principle be empirical evidence concerning the instantiation of supernatural properties (such as *being a god*). If there were such empirical evidence, then *being a god* would count as a natural property (and presumably, any god would count as a natural entity). But Sturgeon objects that "It is not plausible that the success of this sort of natural [i.e. empirical] theology would show that the divine attributes were really natural properties" (2006b: 109). While I do not have any knockdown argument against Sturgeon on this point, I do want to enter into the record that I am less troubled by the prospect of "naturalizing" gods and their attributes. If there were empirical confirmation that God exists, I would be inclined to say that the natural world includes one more entity than I had previously supposed.

what it is for one person to treat another as a mere means, I presume that a property such *as treating no one as a mere means* will also come out as natural according to NP.

1.3.2. Metaphysical naturalism.

It is worth considering a somewhat different way of getting at the commitments of ethical naturalism. This can be accomplished by examining the commitments of metaphysical naturalism, of which ethical naturalism is a special case.[19] Consider, then, the following characterization of metaphysical naturalism:

> The task of the naturalistic metaphysician, as I see it, is simply to draw out the metaphysical implications of contemporary science. A metaphysics which goes beyond the commitments of science is simply unsupported by the best available evidence. A metaphysics which does not make commitments as rich as those of our best current scientific theories asks us to narrow the scope of our ontology in ways which will not withstand scrutiny. For the naturalist, there simply is no extrascientific route to metaphysical understanding (Kornblith 1994: 40).

These comments suggest the following methodological principle, which can be taken as characteristic of a metaphysically naturalist philosophical approach: posit all and only those entities that are needed in our best available scientific theories. Since scientific theories are in the business of providing *a posteriori* or empirical explanations of observable phenomena, it seems to me that it would do no harm if we were to restate the naturalist's methodological principle as follows:

> EC: posit the existence of an entity (or a kind of entity) if and only if reference to that (kind of) entity is needed in our best available *a posteriori* explanations of observable phenomena.

[19] This is not to suggest that all ethical naturalists must accept the broader picture of metaphysical naturalism. For example, someone who accepted the existence of supernatural entities like God might nevertheless embrace the claim that moral properties are identical with natural properties. Even so, it is my impression that a good deal of the interest in the ethical naturalist's project stems from a desire to reconcile moral realism with metaphysical naturalism.

EC (the explanatory criterion) is in accord with what some ethical naturalists have explicitly avowed. Peter Railton, for instance, writes:

> What might be called 'the generic stratagem of naturalistic realism' is to postulate a realm of facts in virtue of the contribution they would make to the *a posteriori* explanation of certain features of our experience. For example, an external world is posited to explain the coherence, stability, and intersubjectivity of sense experience. A moral realist who would avail himself of this stratagem must show that the postulation of moral facts similarly can have an explanatory function (Railton 1986: 171f).

The acceptance of EC explains why the metaphysical naturalist who accepts moral realism is eager to argue that moral facts and properties are identical with or constituted by natural facts and properties: if it is not possible to discover whether or not an individual instantiates a moral property by solely empirical means, then it is hard to see how such a property (or the fact that consists in its being instantiated) would be needed in the best available *a posteriori* explanations of observable phenomena. But if moral properties were not needed in our best *a posteriori* explanations, then, by EC, the metaphysical naturalist ought to deny that there are any moral facts and, thus, he ought to reject moral realism.

1.4. Analytic Ethical Naturalism.

1.4.1. <u>The general strategy of analytic ethical naturalism.</u>

The ethical naturalist's task is to achieve a *naturalistic accommodation* of moral properties and facts. This involves, above all, making the case for the plausibility of the claim that moral properties such as *moral rightness* and *moral goodness* are identical with (or constituted only by) natural properties. The traditional strategy for achieving accommodation is to argue that moral predicates are synonymous with predicates known

to express natural properties. According to the realist version of this strategy, we should view theories in first-order normative ethics as expressing analytic equivalences. Thus, a theory such as AUh, if true, is true in virtue of the fact that the predicate 'morally right' has the same meaning as the predicate 'maximizes hedonic utility.' Because the two predicates are synonymous, the former expresses the same property that the latter expresses. And because, as most metaethicists will grant, the latter predicate expresses a natural property, it follows that 'morally right' expresses the very same natural property. In this way, a successful demonstration of an analytic equivalence between 'morally right' and a natural predicate such as 'maximizes hedonic utility' suffices for the conclusion that *moral rightness* is identical with (and so, is itself) a natural property, as the ethical naturalist claims. Call any view that employs this strategy of naturalistic accommodation a version of *analytic ethical naturalism* (AEN).[20]

1.4.2. Semantic assumptions of analytic ethical naturalism.

To get a better grasp on AEN, and to better understand the ways in which SEN departs from it, it will be useful to have before us a sketch the semantic assumptions that underwrite the former. I begin with some brief preliminaries.

Call any set of individuals at a possible world an *extension*. Call any function that maps possible worlds to extensions an *intension*. The semantic content of any given predicate is identified with an intension. The semantic content (and hence, intension) of a

[20] Here I am interested in explicating what might be thought of as a "classical" version analytic ethical naturalism that relies on a traditional understanding of meaning and conceptual analysis. Since my goal in discussing AEN is simply to provide some motivation for, and a contrast with, SEN, I will not consider more sophisticated versions of AEN that adopt so-called network analyses and two-dimensional semantics. To see an implementation of these resources in defense of ethical naturalism, see Smith (1994) and Jackson (1998).

predicate determines the contribution that the predicate makes to the truth-conditions of sentences in which that predicate appears. Thus, the sentences 'x is F' and 'x is G' have the same truth-conditions just in case the intension of 'F' is the same as the intension of 'G.'

The semantic theory that grounds AEN is sometimes called *descriptivism*. By this view, the intension (and, hence, semantic content) of a speaker's use of a predicate, 'F,' is determined by a description that she "associates" with 'F.' More precisely, the intension of 'F,' as used by a speaker, S, is the function that maps each possible world to the set of individuals at that world that satisfy the description that S associates with 'F.' Some descriptivists urge that we think of the relevant description not as a linguistic entity, but rather, as a property or collection of properties that speakers associate with a given predicate (Jackson 1998: 203f). In either case, the description associated with a predicate should be thought of as something like a criterion specifying the necessary and sufficient conditions that any given individual must satisfy in order for that predicate to be correctly applied to it. In fact, it may be less misleading in some cases to speak of these criteria as *senses* or *concepts*,[21] rather than as descriptions: we might suppose that a speaker has a concept that she associates with 'red' even if she cannot produce an interesting description that applies to all and only red things (perhaps because her concept is more like a pictorial image than a list of properties). The *meaning* of a predicate as used by a speaker should be identified with the concept or description that she associates with it. Thus, where 'F' and 'G' are non-indexical predicates, an utterance of 'F' is synonymous

[21] As I will be using the word 'concept', the concept associated with a predicates is something that (according to descriptivism) fixes or determines which intension is expressed by that predicate. By my usage, a concept should not be identified with the intension itself; nor should it be identified with the property expressed by the predicate. In this, my usage differs from that of (e.g.) Carnap (1947/1956: 21).

with an utterance of 'G' just in case the concept or description associated with one is the same as the concept or description associated with the other.

Something needs to be said about this three-place relation of "association" that obtains among speakers, concepts, and predicates. If descriptivism is to serve the needs of AEN, then facts about which concept a given speaker associates with a given predicate must be discernible for that speaker by *a priori* introspection. Thus, the relation of association should be understood as something like an introspectively accessible psychological state of a speaker. It is for this reason that the form of descriptivism that underwrites AEN must be construed a form of *semantic internalism*. According to semantic internalism, the semantic content of a token predicate is fixed solely virtue of how things are in the mind of a given speaker; facts about the environment outside of the speaker's mind do not directly contribute to fixing the content of the predicates she utters. It follows from this internalist construal of descriptivism that the matter of which intension and content is expressed by a given predicate 'F' depends entirely upon the introspectively accessible psychological states of the speaker who utters 'F.'[22]

A final point concerns a metaphysical assumption that goes along with the descriptivist semantics that underwrites AEN: for each intension there corresponds one and only one property.[23] Indeed some descriptivists, such as Carnap, *identify* intensions with properties (Carnap 1947/1956: 19). Construing properties this way entails that any two predicates that share the same intension will also express the same property; in other

[22] This accords with the account of intension individuation advanced by Rudolph Carnap. He writes, "the intension of a predicate '*Q*' for a speaker *X*, is the general condition which an object *y* must fulfil in order for *X* to be willing to ascribe the predicate '*Q*' to *y*" (1955/1956: 242)

[23] Note that this assumption will not be shared by an ethical non-naturalist. A non-naturalist who accepts AUh will agree that 'morally right' has the same intension as 'maximizes the balance of pleasure over pain'; but he will deny that the two predicates express the same property.

words, co-intensionality is sufficient (and necessary) for property identity (cf. Carnap *ibid*.: 18f). Moreover, because predicates that share the same meaning also share the same intension, the descriptivist semantics sketched here implies that synonymy of predicates is sufficient to establish property identity.[24] In other words, a speaker's utterance of 'Necessarily, for any x, x is F iff x is G' is true just in case her utterance of 'F' is synonymous with her utterance of 'G'. Indeed, this necessity statement is analytic, since its truth depends solely upon the meanings of the words that compose it.

1.4.3. First-order normative ethics as conceptual analysis.

For the analytic ethical naturalist who accepts both moral realism and the descriptivist semantics sketched above, first-order ethical theorizing is an exercise in conceptual analysis.[25] In order to discover whether (e.g.) AUh is true, we must investigate the description or concept that we associate with the predicate 'morally right'. If it should

[24] Hence, this version of descriptivism implies what Brink calls "the semantic test of properties" (see his 1989: 162).

[25] It should be acknowledged that not every naturalistic analysis of moral predicates has the result that first-order moral theorizing proceeds by way of conceptual analysis. Consider, for example, the following ideal observer analysis: (IO) 'ϕ is morally wrong' =df. 'ϕ is forbidden by the moral standard that would be endorsed of by all observers in suitably idealized epistemic conditions.' Even if we agree that IO is the correct metaethical theory, we still need to engage in further inquiry in order to determine what the correct moral standard is and which actions are morally wrong. Indeed, two people could assent to IO and yet disagree as to whether AUh, CI-2, or some other standard is the correct substantive theory in first-order normative ethics of behavior. To discern which of these theories is correct presumably requires a non-conceptual investigation into what standards an ideal observer would endorse (Firth suggests that this would be an empirical investigation, drawing primarily on the resources of psychology [1952:325ff]).

My own view is that analyses along the line of IO are far more plausible than analytic versions of AUh, CI-2, and the like. The trouble, however, is that IO is incompatible with moral realism: IO identifies *moral wrongness* with (roughly) the property of *being forbidden by the moral standard that would be accepted by all idealized observers*. Because of this, if IO is true, then the matter of which actions are morally wrong depends upon facts about which moral standard the ideal observers accept. This, however, is a clear violation of the stance-independence clause associated with realism. The most robust metaethical view that IO can deliver, then, is a naturalistic version of moral constructivism.

The lesson is that not every naturalistic analysis of moral predicates that has been proposed could be put to service in defense of moral realism. My suspicion is that only those analyses that double as theories in first-order ethics will avoid this blatant violation of stance-independence. (Although, as we will see in §1.5.6, there is reason to worry that *all* forms of analytic naturalism violate stance-independence). I will not pursue this suspicion any further, since my primary interest in Realistic AEN is as a contrast to SEN.

turn out upon analysis that the concept that we associate with 'morally right' is the same concept that we associate with 'maximizes hedonic utility' then we can conclude that the two predicates are co-intensional and that AUh is, in fact, true (and analytic at that). More importantly from a metaethical point of view, we could conclude that 'morally right' expresses the same property as 'maximizes hedonic utility'; and since it is not in dispute among metaethicists that latter predicate expresses a natural property, we may conclude that the property expressed by the former is also natural. In that case, it will have been shown that the property *moral rightness* just is the natural property *maximizing hedonic utility*. In this way, AEN (when armed with a successful naturalistic analysis of all moral predicates) achieves a naturalistic accommodation of moral properties and facts.

1.5. The Rejection of Realistic Analytic Ethical Naturalism.

Let us turn now to considerations that have lead metaethicists, including proponents of SEN, to reject analytic ethical naturalism. My treatment will be brief, since my interest here is only to outline the considerations that motivate the adoption of SEN for proponents of moral realism and ethical naturalism.

1.5.1. The rejection of analyticity.

One motivation for rejecting AEN that deserves brief mention is a general skepticism about the existence of any analytic truths whatsoever. An influential case for this skepticism can be found in Quine's "Two Dogmas of Empiricism" (1951).[26] Among the

[26] Of the proponents of SEN, at least Boyd expresses doubt as to whether there are any interesting analytic truths at all. He writes, "On Quinean grounds, I doubt that…analytic definitions or specifications of necessary and sufficient conditions are ever to be found in the case of philosophically important concepts" (1979: 378f; cf. 1988: 196).

arguments Quine advances, there is this: a statement is analytic only if it is incorrigible or immune to revision; no statement is incorrigible; therefore, no statement is analytic (*ibid*.: 40). Obviously, if no statement is analytic, it follows that, contrary to AEN, no theory in first-order ethics can have the status of an analytic truth.

1.5.2. Doubts about descriptivism.

The classical version of descriptivism that underwrites AEN is vulnerable to several well known objections. I will very briefly mention three. First, speakers are sometimes able to express one property rather than another using a predicate even when the concept that she associates with that predicate does not distinguish between the two properties.[27] For instance, although Kant offers up *yellow metal* as the description that he associates with gold, (Kant 1783/2004: 72f) we take him to speak falsely when, pointing to a sample of iron pyrites (fool's gold), he utters 'this stuff here is gold.' But if descriptivism were true, we would have to say that he speaks truly (Kripke 1980 116-119; cf. Putnam 1975b:226f; Donnellan 1970). Second, speakers are sometimes able to express a particular property using a predicate even though the description they associate with that predicate is erroneous. For example, it is argued that speakers who, because they were ignorant of marine biology, associated *being a fish* with the predicate 'whale' nevertheless spoke truly when, pointing to a humpback whale, they uttered 'this is a whale.' If descriptivism were true, however, this would not be so: the object that the

[27] Although I continue to speak of predicate expressions here, I should acknowledge that most of the attacks on descriptivism have focused on the theory as an account of referring expressions such as names and natural kind terms. For the sake of continuity, I tailor these objections to apply to natural kind predicates. As far as I can see, there isn't any reason to think that arguments suitable for refuting a descriptivist treatment of the names of kinds (e.g. 'the tiger') won't be equally effective to refute a descriptivist treatment of corresponding predicates (e.g. 'tiger').

speaker is pointing to does not fall within the intension that that corresponds to her concept (since that intension would include only individuals at a world that are fish) (Kripke *ibid*.). The third objection to descriptivism is related to the previous one: in cases where speakers come to have new beliefs about the nature of a kind—for instance, that whales are not fish, but mammals, descriptivism seems to entail that the semantic content of their predicate (e.g. 'whale,' or else 'fish') has changed. But this, some have argued, is incorrect (Kripke *ibid*. 138). An additional objection worth mentioning is Hilary Putnam's famous "Twin Earth" argument (1975b: 223ff). (A discussion of this argument can be found in the next chapter).

The objections outlined here have lead a number of philosophers to reject classical descriptivism, at least as it applies to proper names and natural kind terms. It might be argued, of course, that moral predicates are less like natural kind predicates, such as 'gold' and 'tiger,' and more like the predicate 'bachelor' (for which descriptivism is still thought to offer a plausible account). Even so, the objections to descriptivism about natural kind predicates deserve mention since they seem to have played a role in turning philosophers away from AEN.

1.5.3. The open question argument.

I turn now to objections against AEN that fall more narrowly within the purview of metaethics. Arguably the most well known of such objections is the open question argument. The *locus classicus* for the open question argument is G. E. Moore's (1903: 66-69), though different versions of the argument have been advanced by Ayer (1936/1952: 104f) and Hare (1952: 83-93, 154f), among others. Without attempting to be

faithful to Moore's own exposition,[28] here is a brief sketch of (one version of) the open question argument: Consider this question: (Q1) "Is act-token ϕ, which maximizes hedonic utility, morally right?" If 'morally right' is synonymous with 'maximizes hedonic utility,' then any competent speaker of English who fully understands Q1 should know the answer to it.[29] But it seems that there could be (indeed, it seems that there *are*) competent speakers of English who fully understand Q1 but do not know the answer to it. For them, Q1 remains an "open question." It follows that 'morally right' is not synonymous with 'maximizes hedonic utility.' Consequently, when AUh is construed as an analytic definition of 'morally right,' it is false. Proponents of the open question argument claim that this argument generalizes; they maintain that we would arrive at a similar conclusion for any other naturalistic analysis of 'morally right' that might be offered (and similarly for naturalistic analyses of 'morally good' etc.). If they are right, then AEN cannot succeed.[30]

---

[28] For a thorough and sympathetic exposition of Moore's open question argument, see Feldman (2005).
[29] This premise may look too strong at first glance, but consider what we would say about a speaker who failed to know the answer to the question (Q2) 'Is Mr. X, a man who, though eligible for marriage, has never been married, a bachelor?' I think we would be strongly inclined to suppose that this speaker does not understand the meaning of at least one of the words in Q2 and so, we would conclude that she is not a competent speaker with respect to Q2. The same holds, I believe, for the questions (Q3) 'Is Fuzzy, who is a female fox, a vixen?' and (Q4) 'Is Alex, who is a male sibling of Bert, the brother of Bert?'
[30] The standard defense of AEN against the open question argument is to maintain that there might be "unobvious synonymies". In the recent metaethical literature, this line of reply has been pressed by Jackson (1998: 151) and Smith (1994: 37).

1.5.4. Chauvinistic conceptual relativism.[31]

Another objection to AEN is that it entails a kind of conceptual relativism that makes genuine moral disagreement impossible between speakers (or linguistic communities) that subscribe to different moral standards (Blackburn 1984: 168; 1998: 14f; Boyd 1988: 186f; Gibbard 1990: 9-18; Hare 1952: 49,148f; Moore 1903: 62ff; Sturgeon 1984: 327f).[32] To illustrate, let's suppose there are two islands that had been colonized by English speakers in the 17th century, but which have not had extensive contact with the outside world since. Suppose further that the denizens of each island are homogenous with respect to the moral code that its members subscribe to. On the Island of Benthamania, (nearly) all denizens accept AUh. Among other things, they tend to feel anger and resentment towards those (and only towards those) who knowingly perform actions that fail to maximize hedonic utility. On the island of New Immanuel, (nearly) all denizens accept CI-2. Among other things, they tend to feel anger and resentment towards those (and only towards those) who knowingly perform actions that treat others as a mere means. If, as the realist construal of AEN requires, we view AUh and CI-2 as expressing analytic definitions of 'morally right,' then we must conclude that in this scenario that 'morally right' has a different meaning in the mouths of Benthamanians than it does in the mouths of New Immanuelers. In addition, we must conclude that

[31] I borrow the phrase 'chauvinistic conceptual relativism' from Horgan and Timmons (1996a: 15). For them, a relativistic construal of a given class of predicates is chauvinistic if "it entails lack of genuine disagreement in cases where two speakers utter apparently contradictory statements which really *are* contradictory." Perhaps they chose the term 'chauvinistic' to describe this phenomenon because a further apparent implication of this sort of relativism is that groups of speakers who do not accept the sort of first-order normative theory that we accept simply fail to have any moral vocabulary at all. What is chauvinistic, then, is that this relativism implies that only we have a moral vocabulary, while all those who fail to share our moral values simply lack one.

[32] Although this objection is articulated earlier by G. E. Moore, R. M. Hare's (1952), with its memorable example involving a missionary among cannibals, is more widely cited and appears to have the status of the *locus classicus* for this particular argument.

Benthamanian and New Immanueler uses of 'morally right' express different intensions and properties. In the mouth of a Benthamanian, 'morally right' expresses the property *maximizing hedonic utility*. In the mouth of a New Immanueler, 'morally right' expresses the property *treating no one as a mere means*. It follows that Benthamanian uses of 'morally right' and New Immanueler uses of 'morally right' contribute differently to the truth-conditions of sentences in which they appear. Thus, when Benthamanians utter 'organ harvesting is morally right' and New Immanuelers utter 'organ harvesting is not morally right' they are not engaging in a substantive moral disagreement; because their uses of 'morally right' are incommensurable with one another, the two parties are merely talking past each other. But this seems incorrect. It is obvious that the Benthamanians and the New Immanuelers are having a substantive moral disagreement and are not talking past each other. If so, then AEN should be rejected; it (at any rate, the semantics required to sustain it) commits us to an implausible kind of conceptual relativism with respect to moral predicates.[33]

---

[33] My presentation of the argument here is directed against those versions of AEN that attempt to analyze 'morally right' in terms of putative right-making properties. It might be thought that a more "indirect" version of AEN—one that offers, for example, an ideal observer analysis of 'right'—would be more successful. By this proposal, we should suppose that both Benthamanians and New Immanuelers associate the property of *being permitted by the moral standard that all ideal observers would endorse* with their use of 'morally right'. In this way, they express the same intension and property with their uses of 'morally right.' The disagreement between the two parties concerns the matter of which moral standard all ideal observers would endorse. But this is a substantive (and potentially empirical) disagreement.

There are at least two reasons why this strategy will not work in the present context. First, as we saw in note 25, ideal observer accounts of moral properties are not compatible with moral realism. Since I am here interested in AEN as a way of preserving moral realism in the face of a commitment to metaphysical naturalism, the present suggestion is of no use. Second, there are doubts about whether an ideal observer can be sufficiently described so that we have reason to think that there is a determinate fact about which moral standard she would endorse. If there is not a determinate fact about this, then there may be excessive moral indeterminacy. But even if the ideal observer can be described in enough detail, a new worry arises: two communities of speakers may associate different descriptions with their conception of an ideal observer. If so, then the content of 'right' in each community will again express different properties and, thus, chauvinistic relativism returns.

There may be, of course, other indirect versions of AEN (ones that do not directly build a moral standard into the content of 'morally right') that better suit the needs of moral realists. I leave it to proponents of AEN to produce such an account. My hope is that I have said enough here to make clear that

1.5.5. The argument from normativity.

The next kind of argument that I want to consider is has been directed against moral realism in all of its varieties. I am of the mind, however, that this sort of objection is especially problematic for AEN since, in my estimation, AEN has fewer resources available to answer it than do other forms of realism (such as ethical non-naturalism and SEN). The argument comes in two varieties. One focuses on a supposed link between moral judgment and motivation. The other focuses on a supposed link between moral facts and reasons for action. I consider each argument in turn.

The argument from motivation, which has its roots in Hume's writings, begins with an assertion of *moral judgment internalism*, the thesis that there is a necessary connection between judging that an act is obligatory for oneself and being motivated to perform that act. More specifically,

> MJI: Necessarily, if S judges (*de se*) that S is morally obligated to ϕ, then S is *pro tanto* motivated to ϕ.[34]

MJI is typically asserted as an analytic, conceptual truth. The trouble it raises for moral realists is that, by the standard theory of motivation (which is, again, associated with Hume), an agent is motivated to ϕ only if she has some desire that would be served by ϕ-ing. If this Humean account of motivation is correct, then it follows from MJI that an agent judges that she is morally obligated to ϕ only if she has some desire that would be satisfied by her ϕ-ing. But it is hard to see why we should think that a speaker's making

---

chauvinistic conceptual relativism presents enough of a challenge to realistic analytic naturalism to produce some motivation for seeking a form of naturalism that does not construe sentences that express necessary co-variance between moral and natural properties as analytic.

[34] MJI, or something very much like it, has been endorsed by Hume (1739/2000: 294-299), Stevenson (1937: 16), Hare (1952: 172), Nagel (1970: Chh. 1, 2), Harman (1975), Mackie (1977: 40), Blackburn (1984: 188), Korsgaard (1986), Dancy (1993: Ch. 1), and Smith (1994: ch. 3), among others.

a moral judgment entail the existence of a desire in this way unless moral judgments are themselves expressions of speakers' desires;[35] but to allow that moral judgments express conative states such as desires is to abandon moral cognitivism. In that case, there would be no point in maintaining moral realism.

One way to answer the anti-realist argument from moral internalism is to reject the Humean theory of motivation. This maneuver requires belief in the existence of "intrinsically motivating propositions." For reasons that are not always well-articulated, naturalists have generally been reluctant to dispense with the Humean theory of motivation and have agreed with Mackie's contention that intrinsically motivating facts would be a "queer sort" of entity (Mackie 1977: 40f; cf. Brink 1997: 12-15; Sturgeon 1992: 100f; 2002: 195). Instead, the standard response to the argument from motivation is to deny MJI. But if proponents of MJI are correct in supposing that the thesis is a conceptual truth, it is hard to see how analytic ethical naturalists can reject it without entering into a stalemate over conceptual intuitions.

The other version of the normativity argument takes as its starting point a thesis that is sometimes called *moral rationalism*:[36]

> MR: Necessarily, if S is morally obligated to ϕ, then S has a *pro tanto* normative reason to ϕ.[37]

---

[35] There are other alternatives that merit acknowledgement. For one, MJI can be satisfied by adopting a cognitivist-subjectivist analysis of *moral obligation*. Such an analysis might look something like this: 'ϕ is morally obligatory =df. 'I disapprove of the failure to ϕ.' But this subjectivist maneuver does the realist no good; it violates realism's stance-independence clause and so represents a constructivist account of moral facts.

[36] In Brink's favored taxonomy, 'agent internalism about reasons' denotes what I am calling 'moral rationalism' (1989: 40). I prefer the latter name since it cuts down on the number of philosophical theses that go by the name 'internalism'.

[37] MR, or something like it, is endorsed by Dancy (1993: 4), Harman (1975: 8), Mackie (1977: 29), Joyce (2001: ch. 2), Shafer-Landau (2003: ch. 8) and Smith (1994: 182ff), among others.

As with MJI, MR is claimed to be an analytic, conceptual truth. It raises trouble for naturalistic moral realism because, according to the standard naturalistic account of normative reasons, an agent's normative reasons are relativized to her desires. When this account of reasons is combined with MR, it follows that an agent's moral obligations are also relativized to her desires. But this is implausible. Surely, a very powerful psychopath is morally obligated to refrain from torturing his victims even if this serves none of his desires. In order to accommodate MR and a plausible first-order account of our moral obligations, it looks like we must reject the thesis that normative reasons are desire-relative. But this raises some dangers for AEN since there is reason to think that desire-independent normative reasons (sometimes called "external reasons") cannot be accommodated within a naturalistic framework (see Joyce 2001 and §4.5.2 of this dissertation).

1.5.6. Analyticity and stance-independence.

A final problem for AEN that I will mention has not received much attention, though it is hinted at by Nicholas Sturgeon (1986b: 117). This problem concerns the question of whether the conception of moral epistemology that goes along with AEN is compatible with the moral realist's claim that moral facts are stance-independent. For a proponent of AEN, the question of which first-order moral theory is correct depends upon which description (or which set of properties) we happen to associate with moral predicates like 'right' and 'good.' But it looks as though the matter of which description we associate with 'right' and 'good' just depends upon which moral theory (construed now as a moral standard) we happen to (perhaps tacitly) accept. (In other words: I associate *maximizing*

hedonic utility with 'morally right' if and only if I accept AUh). If so, then the moral facts turn out to be stance-dependent: what makes a theory like AUh true (if it is true) is that we happen to accept AUh as our moral standard. If we had accepted a different theory, such as CI-2, then it would be the case that we associate *treating no one as a mere means* with 'morally right.' But in that case, CI-2 would be the true theory in the normative ethics of behavior instead of (say) AUh. In this way, AEN looks to be incompatible with a genuinely realist construal of moral facts and moral theory.

1.6. Synthetic Ethical Naturalism.

1.6.1. The necessary *a posteriori*.

The analytic ethical naturalist achieves naturalistic accommodation by construing sentences expressing necessities between moral and natural properties as expressions of analytic, *a priori* necessities. As we have seen, this construal of moral theorizing leads to a host of problems. Fortunately for ethical naturalists, advances in the philosophy of language during the 1970's opened up new metaethical possibilities. Perhaps the most important advance was the recognition of property identities and relations of necessity that are *a posteriori* and synthetic (Kripke 1980; Putnam 1975b). Perhaps the most common example cited of an *a posteriori* identity is the claim that water is identical with H_2O. By the necessity of identity, this identity claim entails that, necessarily, something is a quantity of water just in case it is a quantity of H_2O. This latter proposition (like the former) is thought to be synthetic because 'water' is not synonymous with 'H_2O': the description a competent speaker associates with 'water' may be entirely distinct from the

description she associates with 'H_2O'.[38] The proposition is thought to be *a posteriori* because its truth is (and was) discovered, not by conceptual analysis or by synthetic, *a priori* intuition, but by experimental chemistry, which is an *a posteriori* form of inquiry.

Taking the theoretical identities of chemistry as their inspiration, ethical naturalists argue that sentences used to assert identities (or constitution relations) between moral and natural properties should be understood as expressing synthetic, *a posteriori* necessities:

> The [ethical] naturalist can concede that there are neither synonymies nor meaning implications between moral and nonmoral, for instance, natural, terms and still maintain that moral facts and properties are identical with, or constituted by, natural and social scientific facts and properties. The naturalist's identity or constitution claims can be construed as expressing synthetic [*a posteriori*][39] moral necessities (Brink 1989: 166, 175; cf. Boyd 1988; Lycan 1988; Railton 1989: 157; Sturgeon 1985a: 75n16; 1985b: 25f).[40]

This claim, as I see it, is the distinguishing feature of SEN as a form of naturalistic moral realism.

---

[38] The idea is that the description a competent speaker associates with 'water' may include only the superficial qualities of water. For instance, such a speaker might associate with the predicate 'water' the property of *being a clear, potable liquid (at room temperature) that fills the lakes, rivers and oceans of Earth*. By contrast, a competent user of 'H_2O' is likely to associate with it something like the following description: *being a substance composed of molecules consisting in two hydrogen atoms bonded to one oxygen atom*.

[39] Although he does not say so in the passage cited here, Brink later adds (in 1989: 175) that the epistemological status of these moral necessities is *a posteriori*.

[40] To my knowledge the first person to suggest this view of moral identities is Hilary Putnam in his (1975a)—though his remarks are very brief and inchoate. Putnam gives this view a more explicit endorsement later on (1981: 205-208); but by that time he has already abandoned robust metaphysical realism. Thus, it is not clear whether he should be counted as a proponent of SEN since SEN, as I am understanding it, includes a commitment to realism about moral facts. Two other philosophers deserve mention as having proposed early on that moral identities are *a posteriori*, although they do not fall within the SEN camp. The first is Robert Adams. In his (1979: 76), he writes "ethical wrongness *is* (i..e. is identical with) the property of being contrary to the commands of a loving God. I regard this as a metaphysically necessary, but not an analytic or *a priori* truth." It is not clear that Adams should be classed as a proponent of SEN, however, since he identifies *wrongness* with a supernatural property. (On the other hand, Adams suggests both that God plays a causal role in the world and that we can discover the correct moral theory via *a posteriori* inquiry. As I have mentioned above, I am inclined to think that, if there is a god who plays a causal role in the world, and if there is empirical evidence of his presence, then we should count that god as a natural entity.) The second philosopher deserving mention is Gilbert Harman. In his (1977: 19f), he suggests that an ethical naturalist could argue that moral identities are *a posteriori* in order to answer Moore's open question argument. Since Harman rejects moral realism in favor of a relativistic moral constructivism, I do not classify him as a proponent of SEN.

SEN proponents maintain that the question of which natural properties are to be identified with moral properties (or taken to constitute them) is to be settled by substantive, first-order ethical theorizing (Brink 1989: 177f, 238). Sturgeon, for instance, writes that

> "If hedonistic act utilitarianism…turns out to be true, for example, then we can define the good as pleasure and the absence of pain, and a right action as one that produces at least as much good as any other…" (Sturgeon 1985a: 61; cf. Brink; 2001: 162; Railton 1989: 167).

From these remarks it should be clear that synthetic ethical naturalists view first-order theories like AUh and CI-2, which identify the natural supervenience base of moral facts and properties, as expressing definitions of moral terms. In addition, they view these theories as expressing property identities: "The ethical naturalist claims that moral facts and properties supervene upon natural facts and properties, because moral facts and properties *are* natural facts and properties" (Brink 1989: 176, emphasis in the original).[41]

1.6.2. <u>The semantic and metaphysical underpinnings of the necessary *a posteriori*.</u>

In §1.4.2 I noted that AEN rests upon a semantic foundation: *viz*., descriptivism of an internalist sort. SEN also rests upon a semantic foundation: proponents of SEN support their view by adopting an externalist semantics for moral predicates and property names.

According to semantic externalism, the semantic contents of certain predicates are individuated in part by features of the speakers' environment. This semantic view goes hand in hand with the metaphysical doctrine that there exist natural kinds—kinds whose membership is delimited by a "real" (as opposed to "nominal") essence (Boyd 1988: 194-

[41] Brink goes on to clarify that the appearance of 'are' here may represent either the 'is' of identity or the 'is' of constitution (*ibid*.). Presumably a moral constitution claim would look like a one-way conditional of the form "Necessarily, for any ϕ, if ϕ has N in a circumstance of type C, then ϕ is morally right" where 'N' expresses a natural property.

199; Brink 2001: 160ff; Sturgeon 1985b: 26). A kind's real essence is typically understood to be a property (or collection of properties) that is causally responsible for the observable similarities among its members. The significance of natural kinds for semantic externalism is this: Suppose there is a natural kind, K, with real essence, *E*. If there is a predicate, 'F,' that expresses the property of *being a K*, then, for any world, w, and any object, o, o is in the extension of 'F' at w iff o instantiates *E* at w. The resulting function that maps possible worlds to extensions is the intension and semantic content of 'F.' Importantly, competent users of 'F' may have no *a priori* access to the matter of which properties are included in *E*. That is a fact about their environment; it is not settled by how things are in the minds of speakers. As a result of this, speakers will have no purely *a priori* route towards knowing precisely which intension is expressed by their uses of 'F.'

We can now see the connection between semantic externalism, natural kinds, and *a posteriori* necessity. Suppose that there is a predicate 'G' that expresses the property *E*. Given the stipulations above, 'G' is co-intensional with 'F.' Even so, it may be that 'F' and 'G' are not analytically equivalent.[42] In that case, speakers will be unable to know that 'F' and 'G' are co-intensional without the benefit of an *a posteriori* investigation into which properties belong to the real essence of K. It follows that the necessity claim, 'Necessarily, for any x, x is F iff x is G,' is knowable only *a posteriori*.[43]

[42] We can suppose this only if we deny that that the meaning of a predicate is to be identified with its intension. Brink rejects the identification of meaning with intension in his (1989: Ch. 6).
[43] Again, this necessity claim can be construed further as expressing an *a posteriori* property identity if we assume that co-intensionality is sufficient for property identity.

1.6.3. Causal theories of reference.

Semantic externalism needs to be supplemented by an account that specifies how a given real essence comes to fix the intension of a given predicate. While it is possible to adopt a more sophisticated version of descriptivism to accommodate externalism about semantic content, externalists have traditionally adopted some kind of causal theory of reference (or, better, content-fixing).[44] On a view of this sort, the intension of a predicate is fixed by the real essence of the kind or property that bears the right causal relation to speakers' use of that predicate (Kripke 1980; Putnam 1975b). The version of the causal theory of content most closely associated with SEN is Boyd's "causal regulation" account. In the most widely cited formulation of this view, Boyd writes,

> Roughly, and for nondegenerate 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* […]. 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 relatively 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. […]" (Boyd 1988: 195; cf. Boyd 2003a: 538).

Terence Horgan and Mark Timmons have introduced a useful shorthand to denote the right-hand side of Boyd's biconditional. Following them, let us say that when the right hand side the biconditional is satisfied, *k causally regulates* the use of *t*.

Boyd's statement of his theory of reference is not as lucid as one might like. In the remainder of this section, I want to indicate how I understand his account of reference-fixing (and content-fixing) to work. To begin with, I should note one complication. Hitherto, I have been speaking about the relationship between predicates and the semantic contents (i.e. intensions) that they express. However, Boyd's account of

[44] For sketches of sophisticated descriptivism, see Chalmers (1996), Jackson (1998) and Lewis (1970).

reference is stated in terms of singular terms and the kinds that they putatively refer to. To facilitate discussion, I make the following assumption: if a property causally regulates the use of a property-name (e.g., 'being gold', 'tigerhood', 'goodness', etc.), then that property not only serves as the referent of that property name, but also fixes the intension of any corresponding predicate of which the property name is a nominalization (e.g., 'gold', 'tiger', 'good', etc.). In saying that a property (or essence), P, fixes the intension of a predicate 'F', I mean simply that the intension of 'F' is a function that maps each possible world to the set of individuals at that world that instantiate P (or to the empty set, if nothing at the world instantiates P). Although it may be best for the purposes of SEN not to identify properties with intensions, when a property fixes the intension of a predicate, I will say that that predicate expresses that property.[45] In addition, I will assume that when a property as causally regulates the use of a property-name, it also regulates the use of that name's corresponding predicate.

Here, then, is a (very rough) illustration of Boyd's causal regulation theory of reference as I understand it: Let's assume that it is a necessary truth that something is a quantity of (pure) gold iff it is a quantity of substance composed solely of atoms with atomic number 79 (*Au*, for short). Consider, now, the use of 'gold' by speakers living prior to the rise of atomic chemistry. How does Boyd's theory of reference explain the fact that these speakers used 'gold' to express the property of *being Au*, even though they were not in a position to know or believe anything about atomic numbers, etc.? Here is the sort of story Boyd might tell: First, there is a tendency for it to be true that the pieces

[45] Although it is customary in semantics to identify each intension with a property, this may conflict with some commitments expressed by SEN proponents. In particular, Brink appears to favor a sparse conception of properties whereby not every meaningful or contentful predicate expresses a genuine property (Brink 1989: 158f; cf. Armstrong 1978: 19-29).

of stuff to which the predicate 'gold' is applied by these speakers have the property of *being Au*. Similarly, there is a tendency for it to be true that, when *other* properties are ascribed to the pieces of stuff to which 'gold' is applied, those pieces really have those other properties. For instance, when the speakers utter 'everything that is gold is malleable,' it is true of everything that has the property of *being Au* that it also has the property of *being malleable*. I take this to be what Boyd has in mind when he requires for reference that "what is predicated of the term t will be approximately true of k." Second, this tendency for the speakers' ascriptions to be correct is the result of the sorts of causal mechanisms that Boyd describes. I take it that, at bottom, this involves the claim that the dominant,[46] ultimate[47] causal source of these speakers' 'gold'-related beliefs is *being Au* (or its instances).[48] Under roughly[49] these conditions, according to

---

[46] Here I am drawing on Gareth Evan's (1973) for help. The importance of the dominance condition is this: it may be that some number of our 'gold'-related beliefs have their causal origin not in the property of *being Au*, but rather, in *being* FeS_2 (i.e., *being fool's gold*). This can happen when someone comes to believe that 'there is gold in them hills' expresses a truth as a result of having seen instances of *being* FeS_2 in those hills. Nevertheless, provided that *being Au* is the dominant source of speakers' 'gold'-related beliefs (and that *being* FeS_2 is the source of relatively few of those beliefs) we should judge that the property expressed by 'gold' is *being Au*, and not *being* FeS_2 (and not a disjunction of the two properties).

[47] I say 'ultimate' because it will likely be the case that many speakers never have direct contact with the relevant property expressed by their predicate. For instance, it may be that some users of 'gold' have never been in its presence, and that all of their 'gold'-related beliefs are acquired second hand, by testimony from other speakers. In that case, what is important for reference, as I understand Boyd views, is that at the beginning of the chain of speakers from whence the information related to (e.g.) 'gold' came, there is some speaker for whom *being Au* is the dominant causal source of her 'gold'-related beliefs.

[48] Compare this with Boyd's claim that "the sorts of causal connections which are relevant to reference are just those which are involved in the reliable regulation of belief…" (1988: 195).

[49] Why only "roughly"? I add this hedge because it is doubtful that the conditions Boyd lays out are really sufficient to deliver determinate referents or semantic contents to the relevant terms and predicates in all circumstances. The trouble is this: It could turn out that every piece of Au that the speakers have ever causally interacted with was impure and mixed with some other element, for example, oxygen. In that case, there is another property besides *being Au* that is a candidate for the honors of being the property expressed by 'gold': *viz*., the property of *being a compound of Au and O*. This is one manifestation of the so-called "qua problem," noted by Kim Sterelny (1983), among others. Boyd addresses the qua problem (though not by that name) in his (1999c: 58f) and (2003a: 536f). His discussion is difficult and I am sure I do not fully understand his solution to the problem. As best as I can tell, it involves adding an additional condition for reference: "[In] order for t to refer to p, the epistemic access which uses of t affords speakers to the real properties of p must (help to) *explain* the theoretical and/or practical successes achieved in the domains of inquiry or of practice to which t-talk is central" (2003a: 515). Returning to the example involving 'gold,' this means, presumably, that we should expect that the causal connection between our

Boyd's theory of reference, we can say that *being Au* is the real essence that fixes the intension of 'gold' (as used by our sample speakers). The intension of 'gold' among these speakers then, is a function that maps each possible world to the set of all individuals at that world that exemplify *being Au*, if there are any such individuals.

Let us turn, now, to Boyd's theory of reference as it applies to moral terms and examine the way in which it supports the claim that moral necessities are synthetic *a posteriori*.

Boyd proposes that, among English speakers, the use of moral terms like 'moral goodness' and 'moral rightness' is causally regulated by certain natural properties. I understand this to imply that those natural properties also causally regulate the corresponding predicates 'morally good' and 'morally right.' If Boyd's hypothesis is correct, and if his causal regulation account of content-fixing is true, it follows that, for any natural property, *N*, that uniquely causally regulates (e.g.) 'morally right,' the following necessity claim with be true (where 'N' is a non-moral predicate that expresses *N*): 'Necessarily, for any x, x is morally right iff x has N.' As long as *N* is not among the properties that are analytically associated with 'right,' this necessity claim is synthetic. Since it is presumably an empirical question which natural property causally regulates our use of 'right'—and not something that can be discovered by conceptual analysis or synthetic *a priori* intuition—whether or not this necessity claim is true can be discovered only by *a posteriori* inquiry. In this way, semantic externalism, supplemented with a

---

'gold'-related beliefs and *being Au* helps to explain the theoretical and/or practical success achieved through the use of the predicate 'gold,' whereas the connection between our 'gold'-related beliefs and *being a compound of Au and O* does not explain this (or else explains it less well). Among the things that are unclear here is the matter of what constitutes theoretical and practical success of 'gold'-talk among the speakers living prior to modern chemistry. Although I must set aside these worries about the qua problem, I must nevertheless acknowledge that how these matters are resolved by Boyd have the potential to affect the cogency of arguments I will be advancing in later chapters.

casual theory of reference, explains how it is possible for there to be moral identity claims whose truth-value is discoverable only *a posteriori*.

1.6.4. The epistemological commitments of SEN.

It may be worth making note of some further aspects of the epistemological commitments and preferences associated with SEN and the philosophers who defend it. In addition to holding moral knowledge to be *a posteriori*, the principal defenders of SEN endorse a coherentist account of epistemic justification for moral (and non-moral) beliefs (Boyd 1988; Brink 1989: ch. 5; Sturgeon 2002). According to Brink's characterization, moral coherentism "…holds that one's moral belief *p* is justified insofar as *p* is part of a coherent system of beliefs, both moral and nonmoral, and *p*'s coherence at least partially explains why one holds *p*" (1989: 103). On such a view, no moral belief is self-evident or self-justifying. (That is, no moral belief is justified independently of its inferential relations to other beliefs held by the epistemic agent).

Brink (*ibid*. 104), Boyd (1988: 207) and Sturgeon (2006b: 105) all cite the method of wide reflective equilibrium as their preferred characterization of a coherentist method of moral inquiry.[50] According to this method, roughly, an epistemic agent begins with a stock of considered moral judgments, general moral principles, and non-moral background beliefs. She then makes modifications to all three elements with the goal of producing a new set of judgments, principles, and background beliefs that exhibits maximal coherence. (Note that none of the three elements are thought of as incorrigible or as playing a privileged "foundational" role; all three are susceptible to revision in the

[50] The method of wide reflective equilibrium for moral inquiry was articulated by Rawls (1951; 1971/1999; 1980). For an especially clear description of this method, see Daniels (1979).

interest of coherence.) When her moral judgments, moral principles, and background beliefs exhibit maximal coherence, the agent's moral beliefs can be said to be in "reflective equilibrium." In this state, her moral beliefs (taken as a complete system) enjoy maximal epistemic justification for her.

It is expected that, in pursuing reflective equilibrium among her moral (and non-moral) beliefs, an epistemic agent will make use of her moral intuitions. We may understand an agent S to have a intuition that p just in case it seems to S that p, where this seeming is neither (i) the product of S's sensory perception nor (ii) the product of S's introspecting her own mental operations nor (iii) a process of reasoning from other premises held by S that can be introspectively accessed by S (cf. Bealer 2000: 3f; Huemer 2005: 101f). Following Huemer, we can say that a moral intuition is an intuition with a moral proposition as its content (Huemer *ibid*.).[51]

The role of intuitions here raises some questions about the empirical purity of moral knowledge as it is conceived by proponents of SEN. At first sight anyway, moral intuitions have the appearance of being *a priori*. Synthetic ethical naturalists, however, cannot allow this. They cannot allow that moral intuitions are analytic *a priori*, since this would jeopardize their contention that the moral theories arrived on the basis of such intuitions are synthetic. Nor can they allow that moral intuitions are synthetic *a priori*. This is because the synthetic *a priori* is not an empirical way of knowing. In light of the characterization of *natural propertyhood* in §1.3.1, we could not regard moral properties

[51] In formulating this characterization of a moral intuition I draw on the work of Bealer and Huemer. I depart from them, however, in allowing that it is possible that S has a moral intuition that p even if p is the product of a process of reasoning. I require only that any such process of reasoning is hidden from the epistemic agent (i.e., that the process is not accessible to her via introspection). My reason for loosening the account in this way is to avoid begging the question against the account of intuition preferred by naturalists.

as natural properties if synthetic *a priori* intuition is required to know which things instantiate them.

What is needed by the synthetic ethical naturalist, then, is a naturalistically acceptable account of synthetic moral intuition. To fill this need, SEN proponents contend that moral intuitions are really covert inferences from tacitly held background moral theories. They maintain that moral intuitions resemble scientific intuitions in this respect. With respect to intuitions of the latter kind, Boyd writes,

> …[I]t seems overwhelmingly likely that scientific intuitions should be thought of as trained judgments which resemble perceptual judgments in not involving…explicit inferences, but which resemble explicit inferences in science in depending for their reliability upon the relevant approximate truth of the explicit theories which help to determine them (1988: 193).

Sturgeon adds,

> …I find it more plausible to think that what is [in scientific practice] called intuition is actually a product of inference in a broad but epistemologically well-motivated sense. Judgment here outruns the ability to articulate reasons; but we need some account of why the only people with physical intuition worth trusting are those with extensive knowledge of highly sophisticated, approximately true physical theory and a lot of experience in applying it (2002: 203).

Although Boyd and Sturgeon here suggest that the scientific theories that ground scientific intuition are themselves explicitly known, other naturalists note that scientific intuitions can be grounded in tacit theories that epistemic agents cannot easily articulate (Kornblith 2002: 13). Most importantly, all of these naturalists suppose that the relevant background scientific theories are empirically justifiable (if justifiable at all). As a result, the scientific intuitions that are covert inferences from such theories should not be thought of as *a priori*; they are, instead, *a posteriori*, since they are inferences drawn from empirically or *a posteriori* justifiable background theories.

SEN proponents argue that this conception of scientific intuition extends to moral intuitions as well. Thus, Boyd writes

> …[W]e may now treat moral intuitions exactly on a par with scientific intuitions, as a species of trained judgment. Such intuitions are not assigned a foundational role in moral inquiry…[they] are simply one cognitive manifestation of our moral understanding, just as physical intuitions, say, are a cognitive manifestation of physicists' understanding of their subject matter (1988: 207f; cf. Sturgeon 2002: 203).

The idea, as I understand it, is this: When contemplating possible cases, an epistemic agent will often have moral intuitions. Although these intuitions may be "phenomenologically basic," in the sense that "their inferential heritage is not introspectively available" to the agent (Kornblith 2002: 20), they are, nevertheless, the result of tacit inferences from the agent's background moral beliefs and theory. The epistemic agent who has these intuitions can use them as evidence in coherence reasoning, where the goal is to move from the tacit background theory to an explicit moral theory that is more coherent, and thus, better justified than the initial tacit theory.

Naturally, critics of SEN will want to ask where the agent's tacit background theory comes from. Sturgeon suggests that such a theory might be innate, though, in his view, the answer to this question "doesn't much matter" (2006a: 254). I disagree: the genealogy of our tacit moral beliefs is of grave metaethical importance. In the first place, it isn't obvious that our tacit theory has a genealogy that is compatible with the commitments of the naturalist's empiricism. An ethical non-naturalist, for instance, might concede that many of our particular moral intuitions are inferences from a tacit background moral theory but insist, nevertheless, that the tacit theory, inchoate as it may be, enters our minds by way of a synthetic *a priori* grasping. While I am not inclined to accept the non-naturalist line suggested here, I do think the naturalist owes us greater

assurances that the tacit theory has a genealogy that is compatible with his empiricism. But secondly, there is to my mind a more serious challenge to SEN than the worry about preserving the empirical purity of moral epistemology. This challenge asks whether there is a plausible story about the origins of our tacit moral theory according to which that theory is a roughly accurate reflection of moral reality. All SEN proponents recognize that, in order for the method of reflective equilibrium to yield moral *knowledge* (rather than merely justified moral belief), it needs to be the case that the background moral beliefs with which we start are "sufficiently near the truth" (Boyd 1988: 201, 207; Brink 1999: 207; Sturgeon 1985a: 67; 2006a: 254f; 2006b: 105f). The present worry is that the most plausible genealogical stories about the sources of our tacit moral beliefs and theories—including stories according to which they are innate—may turn out to undermine our confidence that such theories really are "approximately true." (This worry is expanded in to a full-blown argument against SEN in Chapter 6 of this dissertation).

1.7. How SEN answers the objections to AEN.

In this, the final section of Chapter 1, I want to revisit the objections to AEN discussed in §1.5 and briefly indicate how SEN promises the ethical naturalist an answer to them. Its answers to first two objections require little comment. In the first place, because SEN takes knowledge of moral claims to be synthetic, it does not run afoul of skepticism about analyticity. Second, since SEN is grounded in an externalist moral semantics, it does not require the kind of internalist-descriptivist semantics thought to be refuted by the arguments of Kripke, Putnam, and others.

The externalist moral semantics also supplies the ethical naturalist with a simple response to the open question argument. The open question argument can be used to show that moral properties are distinct from all natural properties only if it is assumed that two predicates express the same property only if they are synonymous. However, recall from §1.6.2 that semantic externalism makes it possible for two predicates to express the same property even when their meaning (thought of as the concepts speakers associate with them) are distinct. The synthetic ethical naturalist can claim that, just as 'water' and 'H_2O' express the same property despite being non-synonymous, so too is it (epistemically) possible for us that 'morally right' and the natural predicate (e.g.) 'maximizes hedonic utility' express the same property even if these predicates are known not to be synonymous (Boyd 1988: 199; Brink 1989: 165; Lycan 1988: 199-202; Railton 1989: 157f; Sturgeon 1985b: 25f; 2003: 533f).

SEN also promises the ethical naturalist a way out of the problem of chauvinistic conceptual relativism. Given the externalist moral semantics favored by SEN, it does not follow from the fact that two speakers accept a different moral standard (and thus,

associate different descriptions with their use of moral predicates) that their uses of 'morally right' must express different semantic contents. It may be that the very same property causally regulates the use of both speakers' predicates even if (at least) one of them has an incomplete or mistaken concept associated with that predicate. Thus, it could be that the property of *treating no one as a mere means* causally regulates the use of 'morally right' among the Benthamanians (who, recall, accept AUh). If so, Benthamanian uses of 'right' express the same content as the New Immanuel uses of 'right.' The Benthamanians are simply mistaken as to the real nature of the property that regulates their own uses of 'right.' (One possible explanation that might be offered for their error is, perhaps, the fact—if it is a fact—that most of the actions that treat no one as a mere means also maximize hedonic utility.)[52] Whatever the case may be, SEN's externalist moral semantics makes it is possible—in certain cases, at least—for speakers who accept different moral standards to engage in a substantive disagreement about what is morally right or morally good (Boyd 1988: 199, 209f; Brink 2001: 163; Sturgeon 1984: 329; 1986b: 124).

SEN proponents respond to the normativity objections by denying both MJI and MR. For an analytic naturalist, this denial can succeed only if he can show that neither MJI nor MR is part of the concept that we associate with 'moral judgment' or 'moral obligation.' While SEN proponents have followed this strategy (thereby embracing what is sometimes called *moral externalism*),[53] it has two disadvantages. In the first place,

---

[52] In advancing this hypothesis, I am merely trying to indicate how an explanation of their error might go. For what it's worth, I have grave doubts that most actions that treat no one as a mere means also maximize hedonic utility.

[53] Brink argues this way in his (1989: ch. 3). He offers as conceptual possibility the existence of an *amoralist*. There are two salient kinds of amoralist, and Brink contends that both are possible. The first amoralist is an agent who judges an action to be obligatory but has no motivation to perform it. The second

while MJI might be thought to be a vulnerable principle, the denial of MR is much harder to swallow. To deny MR is to allow that it is a conceptual possibility that there is some very powerful psychopath who, while morally obligated to refrain from harming his victim, simply has no reason to so refrain. I believe that the concept of a moral obligation does not allow for such a possibility. A second disadvantage for this strategy is that a straight denial of MJI and MR appears to be doomed to end in stalemate at best. Fortunately, the synthetic ethical naturalist's acceptance of an externalist moral semantics makes available a stronger reply. One of the supposed benefits of an externalist semantics is that it makes it possible for speakers to use a predicate to express a property, even if the concept that those speakers associate with that predicate is erroneous. For example, we saw in §1.5.2 that semantic externalism about biological species predicates makes it possible for a community of speakers to use 'whale' to express the *property of being* a whale even if the description they associate with that predicate mistakenly includes the property of *being a fish*. With this in mind, the SEN proponent is in a position to concede that MJI and MR are both part of the concept or description that we associate with the predicate 'morally obligatory' while denying that the real property of *moral obligation* satisfies either conceptual requirement.[54]

The last objection to AEN concerns the question over whether construing moral theories as expressing conceptual or analytic truths leaves enough room for the possibility (characteristic of robust moral realism) that our best moral theory is nevertheless false. SEN seems able to account for this possibility with less difficulty. According to the

---

amoralist is an agent who has an obligation but, because of his preferences, has no reason to fulfill it. See also Boyd (1988: 214ff) and Sturgeon (1986b: 121f).

[54] I am not aware of any synthetic ethical naturalist who explicitly advances this argument in precisely this fashion. Brink comes close when he rejects the status of MJI as a conceptual truth on what he calls "Quinean grounds" in his (1987: 294).

semantics of SEN, the matter of which property is expressed by 'morally right' is not settled merely by the concept we associate with the predicate. Because of this, it appears to be in principle possible that we achieve reflective equilibrium among our beliefs about what is morally right while the moral theory we accept fails to capture the true nature of the natural property that causally regulates our use of 'morally right' (i.e., the property *moral rightness*).[55]

[55] Contrast this with Putnam who, writing against metaphysical realism, argues that "The supposition that even an "ideal" theory (from a pragmatic point of view) might really be false appears to collapse into unintelligibility" (1977: 486). If Putnam is right, then even SEN's central supposition that moral terms pick out kinds or properties with *a posteriori* real definitions will not deliver a robustly realistic metaethic.

CHAPTER 2

MORAL TWIN EARTH VERSUS EXTERNALIST MORAL SEMANTICS

2.1. Introduction.

In the previous chapter, we saw that the realist variety of AEN cannot accommodate the existence of genuine moral disagreement between speakers (or communities of speakers) who subscribe to different moral standards; instead, AEN entails a chauvinistic form of conceptual relativism, wrongly implying that speakers disagreeing with one another about moral matters are only verbally disagreeing. As we also saw, the adoption of SEN is thought to promise a way around this problem for the ethical naturalist. Because the externalist moral semantics that grounds SEN makes it possible for two speakers to express the same content with a predicate even when the concepts they associate with that predicate are distinct, the fact that two speakers associate different moral standards with their uses of 'morally right' need not entail that they express distinct properties; and if their predicates do express the same property, they will be able to use those predicates to engage in substantive moral disagreement, even if they happen to associate different moral standards with their use of those predicates.

In a series of papers, Terence Horgan and Mark Timmons (H&T) have advanced an argument that shows that this supposed advantage for SEN over AEN is illusory: the externalist moral semantics favored by SEN also entails a chauvinistic form of conceptual relativism. If they are right, then the externalist semantics that serves as the foundation of SEN should be rejected.

H&T's argument centers around their "Moral Twin Earth" thought experiment (MTE). In constructing this thought experiment, H&T draw inspiration from the famous Twin Earth thought experiment introduced by Hilary Putnam in his "The Meaning of 'Meaning'" (1975b). However, whereas Putnam's thought experiment convinced many that semantic externalism offers the correct semantics for natural kind terms like 'water' and 'aluminum,' the MTE thought experiment is meant to show that externalist semantics do not offer a correct account of the contents of moral predicates. In the present chapter, I begin by sketching Putnam's original Twin Earth argument for natural kind terms. I then present H&T's MTE thought experiment and the accompanying argument against semantic externalism for moral predicates. In the remaining sections, I consider several challenges to H&T's argument. These challenges all involve the claim that the MTE thought experiment is misleading and that the intuitions generated from it—intuitions that undermine SEN's semantics—should not be trusted. If these objections are well-founded, then we should reject H&T's argument against SEN. I argue, however, that these objections miss their mark: there is no reason to suppose that the MTE thought experiment is misleading. The intuitions generated by the thought experiment are at least as innocent as the intuitions generated by Putnam's original. (In Chapters 3 and 4, I go on to consider other attempts to defend SEN from the MTE argument.)

2.2. Putnam's Twin Earth.

Putnam asks us to imagine a planet in our galaxy that is as near a duplicate to Earth as is possible save for the following difference: on Twin Earth "…the liquid called 'water' is not H_2O but a different liquid whose chemical formula is very long and complicated."

Putnam calls this chemical formula 'XYZ.' Despite its different chemical structure, XYZ is indiscernible from H_2O in all of its superficial properties. In addition, the role that XYZ plays on Twin Earth is very similar to the role that H_2O plays here on Earth: e.g. it fills the lakes, rivers, and oceans; nourishes Twin Earthling life forms; etc. (Putnam 1975b: 223).

In light of the Twin Earth scenario, there are two salient alternative ways of describing the extensions of 'water' as used by Earthlings and 'water' as used by Twin Earthlings. (Hereafter, I use *t-'water'* to represent Twin Earthling uses of 'water.') In the first place, we can say that 'water' and t-'water' have the same extension. This extension includes all molecules of XYZ and all molecules of H_2O and nothing else. By the second alternative, we can say that 'water' has all and only molecules of H_2O as its extension, while t-'water' has all and only molecules of XYZ as its extension. Putnam contends (and much of the philosophical community agrees) that the second alternative is correct.

Putnam recognizes that he has not yet established semantic externalism as an account of the semantic contents of predicates like 'water.' For all that has been said, our judgment that 'water' and t-'water' differ in extension may be due to our supposing that Earthlings and Twin Earthlings have different 'water'-related concepts or beliefs. After all, many present day Earthlings believe that the stuff they call 'water' is composed of H_2O. Given Twin Earth's similarities with Earth, we should expect that many Twin Earthlings believe that the stuff they call t-'water' is composed of XYZ. Thus, it may be that the difference in extension is a result of differences in the internal[1] mental states of

[1] Here I use 'internal' where others might use 'narrow.' An individual's narrow mental states are typically understood to be those mental states that the individual shares with all of her intrinsic duplicates. Some

Earthling and Twin Earthling speakers, and not simply a result of differences in their environments, as Putnam wants to claim.

To block the semantic internalist's preferred explanation of the intuition that 'water' and t-'water' have different extensions, Putnam asks his readers to imagine both planets as they were in 1750, prior to the rise of modern chemistry. It may now be plausibly supposed that the beliefs and concepts that Twin Earthlings associate with t-'water' are exactly like—if not identical to—the beliefs and concepts that Earthlings associate with 'water.' But even in this case, Putnam contends (and many philosophers agree) that "the extension of the term 'water' was just as much H_2O on Earth in 1750 as in 1950; and the extension of the term [t-]'water' was just as much *XYZ* on Twin Earth in 1750 as in 1950" (*ibid*. 224).

The pre-chemistry Twin Earth case shows that it is possible for two speakers A and B to associate the same descriptions or concepts with a kind predicate 'F' and yet A expresses a different intension using 'F' than B does.[2] It follows that the intensions (and hence semantic contents) of these predicates are not fixed solely in virtue of the concepts that speakers associate with them. Something external to the speaker's mind is needed. The upshot of Twin Earth, then, is that semantic externalism offers the correct account of the semantic contents of natural kind predicates like 'water.'

---

have noticed that, on this construal, it isn't clear that an Earthling and her Twin Earth counterpart can have the same narrow mental states. The problem is that over half of the Twin Earthling's body is presumably composed of XYZ molecules, whereas the same proportion of the Earthling's body is composed of H_2O molecules. If so, the two are not intrinsic duplicates, strictly speaking. Thus, they cannot have the same narrow mental states. In light of this, I prefer the somewhat less loaded 'internal mental state.' This locution is meant to capture intuitively whatever kind of mental state Putnam intended each Twin Earthling to share with his or her Earthling counterpart in the 1750 story.

[2] Putnam does not himself state the conclusion of the Twin Earth argument in terms of intensions. This may be because in "The Meaning of 'Meaning'" he is thinking of intensions as "something like an individual speaker's *concept*" (1975b: 245, cf. 216-219). I am not using 'intension' in this way. I take intensions to be functions from worlds to extensions. The intension of a term need not be fixed by speakers' concepts. If intensions are thought of as functions from worlds to extensions, then it follows that if two predicates have different extensions at any single world, they have different intensions.

It is worth observing that causal theories of reference such as Boyd's nicely explain the intuition that 'water' and t-'water' have different extensions. On Earth, speakers have causal contact with H_2O but not with XYZ. On Twin Earth, speakers have causal contact with XYZ but not with H_2O. A defender of a causal theory of reference might argue that the intuition that 'water' and t-'water' have different extensions is due to our tacit recognition that some sort of causal acquaintance is necessary for successful reference.

2.3. Moral Twin Earth.

H&T's Moral Twin Earth thought experiment draws inspiration from Putnam's thought experiment in order to undermine the application of Boyd's semantics to moral terms. We have already met Boyd's causal theory in §1.6.3. Here is H&T's formulation of this theory as it applies to moral terms:

> CSN *Causal semantic naturalism*: Each moral term t rigidly designates the natural property N that uniquely causally regulates the use of t by humans (1990-91: 455).

We should understand CSN to include the claim that each moral predicate, 'F,' expresses the natural property that uniquely causally regulates the use of 'F' by humans.[3]

The Moral Twin Earth thought experiment (MTE) begins with the stipulation that on Earth "human uses of 'good' and 'right' are causally regulated by certain *functional* properties;[4] and that, as a matter of empirical fact, these are consequentialist properties

[3] That is to say, if there is a natural property *N* that uniquely causally regulates the actual use of 'F,' then the intension of 'F' is the function that maps each possible world to an extension containing all and only instances of *N* at that world.

[4] By stipulating that moral properties are functional properties, H&T are following Brink's suggestion in his (1984: 121f). There, Brink invokes a functionalist account of moral properties in order to explain the supervenience of moral properties on physical properties.

whose functional essence is captured by some specific consequentialist normative theory; call this theory T^c" (1990-91: 458).[5] I find that it makes for easier discussion if we name one of these consequentialist properties. Let us add to H&T's stipulations that the property of *maximizing utility* causally regulates the use of 'morally right' among Earthling speakers.

Moral Twin Earth is a planet in our galaxy that is as near a duplicate of Earth as possible save for one difference to be noted shortly. But first, let us highlight an important similarity between the two planets: like us, Twin Earthlings

> …use the terms 'good' and 'bad,' 'right' and wrong' to evaluate actions, persons, institutions and so forth…[T]he terms are used to reason about considerations bearing on Moral Twin Earthling well-being; Moral Twin Earthlings are normally disposed to act in certain ways corresponding to judgments about what is 'good' and 'right;' they normally take considerations about what is 'good' and 'right' to be especially important, even of overriding importance in most cases, in deciding what to do, and so on (1990-91: 459).

Despite this similarity, on Moral Twin Earth the use of these terms is causally regulated by certain functional properties whose essence is captured by a non-consequentialist, deontological normative theory called 'T^d'. Again, I find it useful to name one of these properties. Let's say that the property of *treating no one as a mere means* causally regulates the use of 'morally right' on Moral Twin Earth. (Hereafter, I use *t-'right'* to represent uses of 'morally right' by Twin Earthlings.) To account for this respect in which Twin Earthlings differ from Earthlings, H&T stipulate that there are "species-wide differences in psychological temperament" between Earthlings and Twin Earthlings (*ibid*. For more on this, see §2.5.3 below).

[5] Recall from Chapter §1.6.1 that SEN proponents suppose that the essence of moral properties will be captured by theories in first-order normative ethics (see Brink 1989: 177f, 238; 2001: 162; Sturgeon 1985a: 61).

With this description of the MTE case in hand, H&T present the reader with two alternative ways to describe the uses of 'good' and 'right' by Earthlings and Twin Earthlings. On the first interpretation, "moral and twin-moral terms *differ in meaning*, and are not intertranslatable." On the second interpretation, "moral and twin-moral terms do *not* differ in meaning or reference" (1990-91: 460, emphasis in the original). Although H&T here speak of the sameness of *meaning* as being what is at issue, I think it is better to speak of the sameness of *semantic content*.[6] As mentioned in §1.4.2, I take the semantic content of a predicate to be its intension. The intension of a predicate determines the contribution the predicate makes to the truth-conditions of sentences in which it appears. In my view, then, the question before us is whether we should view Earthling utterances of 'ϕ is right' as having different truth-conditions from Twin Earthling utterances of 'ϕ is [t-]right.' (Putting things this way commits us to moral cognitivism. However, since cognitivism is already a commitment of SEN, I see no harm in assuming it to be true so long as we keep in mind that one possible lesson of MTE is that the content of moral predicates is [primarily] non-cognitive or of some expressivist sort.)

H&T contend that the second interpretation is correct: the content of 'ϕ is right' on Earth is the same as the content of 'ϕ is [t-]right' on Twin Earth. If their judgment is correct—and I think it is—it spells trouble for CSN. Since distinct properties causally

[6] One reason to avoid casting the question as one about meaning is that Brink, for one, separates the meaning of a predicate from the property or intension it expresses. This is apparent in his rejection of "the semantic test of properties" (Brink 1989: 162, 166). He allows that a natural predicate could express the same property as a moral predicate even when the two predicates do not have the same meaning. Brink evidently thinks of the meaning of a predicate as something like a Fregean sense. On such a view, a predicate's meaning is roughly a criterion for its application that speakers associate with that predicate. If this is the way meaning is to be construed, then it is immaterial to the defense of CSN whether 'right' has the same meaning (i.e., associated criterion of application) as t-'right.' For this reason, I find it preferable to pose the present question as being about semantic content rather than meaning.

regulate the use of 'right' and t-'right,' CSN entails that utterances of 'right' and t-'right' express different content. 'Right' expresses the property of *maximizing utility*. T-'right' expresses the property of *treating no one as a mere means*. That these are distinct properties can be seen when we consider organ harvest cases in which an individual person is treated as a mere means in order to maximize utility.[7] Such actions fall within the intension of 'right' but not within the intension of t-'right.' It follows that, if H&T are correct in judging that 'right' and t-'right' have the same content, CSN is false.

The defender of CSN may be tempted simply to reject the judgment that 'right' and t-'right' have the same content. Such a move would commit CSN to a kind of conceptual relativism whereby 'right' and t-'right' are incommensurable. The implausibility of this maneuver is revealed when one considers how things would go if Earthlings and Twin Earthlings were to meet:

> Suppose that Earthlings visit Twin Earth (or vice versa), and both groups come to realize that different natural properties causally regulate their respective uses of 'good,' 'right,' and other moral terms. If CSN were true, then recognition of these differences ought to result in its seeming rather silly, to members of each group, to engage in intergroup debate about goodness—about whether it conforms to normative theory T^c or to T^d. […] But such intergroup debate in the Moral Twin Earth story would surely strike both groups not as silly, but as quite appropriate, because they would regard one another as differing in moral belief and moral theory, not in meaning (Timmons 1999: 62f).

By contrast, in Putnam's original Twin Earth scenario ('PTE,' hereafter) it is plausible to suppose that Earthlings and Twin Earthlings who meet would cease debating about the true chemical composition of "water" as soon as they recognized that different chemical substances causally regulate the use of 'water' and t-'water.' One expects that they would readily acknowledge that their disagreement was merely verbal.

[7] The standard organ harvest case involves a surgeon who kills one innocent healthy patient (without his consent) in order to transplant his organs to five other patients who would otherwise die.

It would appear, then, that MTE refutes CSN. Since CSN is the semantics that underwrites SEN, MTE deals a blow to SEN as well. Nor is there much promise in replacing CSN with an externalist moral semantics that has a different content-fixing mechanism from Boyd's. H&T characterize the basic MTE story as providing a "recipe" for generating arguments against any version of externalist moral semantics that one might propose. They contend that, for any relation that is proposed as sufficient to fix the reference (and, presumably, the contents) of moral terms, one of the following will be true: (a) the relation is insufficient to fix a determinate reference for moral terms; or else (b) a population of Twin Earthlings can be imagined whose moral terms bear the proposed reference relation to a different natural property than Earthling moral terms bear this relation to. In that case, the proposed moral semantics will again result in an objectionably "chauvinistic" form of relativism.[8]

2.4. The Attack on the Moral Twin Earth Thought Experiment.

2.4.1. Introduction.

Several philosophers have argued that H&T's Moral Twin Earth thought experiment is a "flawed-intuition pump." This line of argument receives its most sustained articulation in a jointly authored essay by Stephen Laurence, Eric Margolis and Angus Dawson (LM&D). They claim that the MTE argument provides "no reason at all for rejecting ethical naturalism" (1999: 135). This is because the MTE thought experiment contains misleading features that distort readers' semantic intuitions. To make their case, they highlight the ways in which MTE differs from PTE:

[8] H&T sketch a "generic" form of the MTE argument in their (2000). For a look at MTE in action against other proposed moral semantics see H&T (1996a; 2000; Forthcoming).

> Like H&T, we will not question the legitimacy of Putnam's original Twin Earth thought experiment. Our question is whether H&T's Moral Twin Earth thought experiment is as legitimate as the original. The whole point of H&T's direct argument is that there is supposed to be an asymmetry between the intuitions generated by Twin Earth and Moral Twin Earth; this asymmetry is supposed to argue for the claim that the moral terms aren't rigid designators.[9] For the argument to work, however, the two thought experiments have to be constructed in analogous fashion. The problem with the argument is that they aren't. There are a number of crucial disanalogies between the two thought experiments, and it's these disanalogies that do much of the work in generating the intuitions that H&T's arguments rely upon (1999: 155; cf. Geirsson 2003: 118).

Now, I think LM&D vastly overstate the importance of MTE's connection with Putnam's original. The MTE argument would stand just fine on its own even if no one had ever dreamed up PTE. But even if the connection were important, the MTE argument cannot be undermined simply by pointing to ways in which the MTE thought experiment differs from PTE. Those differences ought to be such that there is good reason to think that they will contribute to the distortion of our intuitions. In the remainder of this chapter, I consider the three disanalogies between MTE and PTE that LM&D highlight. I argue that LM&D fail to show that the respects in which MTE differs from PTE exert (or are likely to exert) a distorting influence on our intuitions. My aim here is to establish only that the MTE thought experiment is no worse an intuition pump than PTE. I have nothing to say in defense of the use of these (or any other) thought experiments more generally.

2.4.2. A preliminary objection.

Before turning to the objections raised by LM&D, I want to deal with a worry about MTE that David Brink raises. In the version of PTE that establishes semantic

[9] Whereas LM&D suggest that what is at stake is whether moral terms are rigid designators, I think the real issue is whether the contents of moral terms are fixed in accordance with the semantic externalist's picture.

externalism (where the story is set in 1750), it is important that the speakers on Twin Earth associate the very same concept with t-'water' that Earthlings associate with 'water.' To ensure that they do, Putnam asks us to imagine that the relevant Twin Earthling speaker in his example ($Oscar_2$) is an exact duplicate of his Earthling counterpart ($Oscar_1$) with respect to his "appearance, feelings, thoughts, interior monologue, etc." (1975b: 224). Since the only relevant difference between Earth and Putnam's Twin Earth is the underlying composition of the "watery" stuff, this supposition can be entertained without much difficulty. But this is not so for MTE. As Brink observes,

> If people have the same commitments to morality on Earth and Moral Twin Earth, the differing standards will cause each planet's people to assess people, actions, and institutions differently; over the long run, this should affect the course of individual and social histories on Earth and Moral Twin Earth. Though the members of both planetary pairs—Earth and Twin Earth and Earth and Moral Twin Earth—are…*otherwise* indistinguishable, this caveat includes many more differences in the second pair than in the first. As it seemed important to Putnam's original arguments that differences between Earth and Twin Earth be minimized, the more extensive differences between Earth and Moral Twin Earth may complicate Timmons and Horgan's argument (2001: 165n21).

Since the inevitable differences in the behaviors of Earthlings and Moral Twin Earthlings are undoubtedly due to differences in their internal mental states, it is hard to see how we can maintain that Moral Twin Earthlings have the very same internal mental states as their Earthling counterparts. Here then, we are confronted with an obvious and glaring difference between MTE and PTE. What should we make of this difference?

What Brink's observation reveals is that MTE could not be used to *confirm* an externalist semantics for moral terms. If, contra H&T, we had judged that 'right' and t-'right' express *different* content, we would not have been free to conclude that the contents of moral terms are (at least partly) individuated by features of the external

environment. For we would not have ruled out the possibility that our judging 'right' and t-'right' to have different content is due to our taking Earthlings and Twin Earthlings to associate different concepts or beliefs with their respective predicates.

This should not worry us. Although MTE could not confirm externalist moral semantics like CSN, it remains an important test that such theories must pass in order to remain viable. In this respect, it is much like Putnam's own initial setup of Twin Earth. Recall that Putnam first warms up his readers with a version of PTE where the relevant users of 'water' and t-'water' are our contemporaries (or nearly so). In that example it is understood that Twin Earthlings have different internal mental states from their Earthling counterparts. Indeed, one such difference is that, whereas contemporary Earthlings believe that 'Water is composed of H_2O' expresses a truth, Twin Earthlings do not believe this. It is important to recognize that it is not trivial that Putnam's audience widely agreed that contemporary Earthlings and Twin Earthlings express different properties using 'water' and t-'water.' If readers had judged otherwise, Putnam's entire argument would have been stopped dead in its tracks. If we had judged that 'water' and t-'water' had the same content in 1950, despite the different concepts and beliefs that speakers on each planet associate with these terms, there would have been no way that our intuitions would be reversed by making Earthlings and Twin Earthlings more alike in their 'water'-related internal mental states. Nor would there be any grounds to complain that the differences allowed in the Twin Earthlings' internal mental states unfairly stack the deck against an externalist semantics for 'water.' If anything, the inclusion of such differences *favors* the sort of intuition that externalists hope to elicit. But if this is so with respect to the 1950 version of PTE, then we must conclude that there are no grounds for

the complaint that MTE stacks the deck against an externalist semantics for 'right' by allowing (indeed, stipulating) that Moral Twin Earthlings have some internal mental states that differ from those of their Earthling counterparts.

2.4.3. First objection: competing theories of a kind.

LM&D observe that in PTE the relevant substance that the Twin Earthlings are causally acquainted with is a fictitious philosophical invention that readers have no familiarity with. They write that XYZ is a "chemical composition that's tied to a chemical theory that no one has ever supposed is true of water" (1999: 156). Although it may be, strictly speaking, an epistemic possibility for us that water is XYZ, it is not a "live" epistemic possibility; it is not a possibility that we take seriously in non-skeptical contexts. To abbreviate this feature of PTE, let's say that PTE *does not involve competing theories of a kind*. In the MTE scenario, by contrast, the properties that causally regulate 'right' and t-'right' *do* answer to competing theories of a kind. Both consequentialism and deontology are live epistemic possibilities for us. As LM&D observe, both theories have "strong advocates in philosophical circles" (*ibid*.).

LM&D argue that this feature of MTE—the fact that it involves competing theories of a kind—distorts readers' intuitions about the case. In particular, it makes it much more tempting for readers to view the parties on Earth and Twin Earth as expressing the same property than would be the case if the Twin Earthlings' moral theory were something with very little plausibility as the correct theory of our own use of 'right.'[10] If it is to be legitimate, the MTE thought experiment should be purged of this

[10] Eric Gample makes the same point in his (1997: 152); see also Merli (2002). Merli argues that, if both Earthlings and Twin Earthlings separately achieve reflective equilibrium with respect to the question of

misleading feature. The story should be retold so that the property regulating t-'right' does not have its essence specified by a moral theory that is a live epistemic possibility for us (*ibid*.).

2.4.4. First reply.

As a matter of fact, I think a compelling MTE thought experiment can be constructed that does not make use of competing moral theories (see §2.4.5 below). For the moment, however, I want to argue that it is by no means obvious that MTE's inclusion of competing moral theories is grounds for criticism. In fact, it seems to me that LM&D's observation poses a greater threat to PTE than it does to MTE. In particular, PTE is vulnerable to the objection that the externalist intuition gets unfair leverage precisely because Putnam's example *fails to involve* competing theories of the nature of water. To see why PTE is vulnerable on this front, consider the following complaint that might be raised by a semantic internalist against Putnam's original argument for externalism:

> Surely, all of Putnam's readers do in fact associate the property *being composed of H_2O molecules* with their use of 'water.' Indeed, many young children know that water is composed of something called 'H_2O' long before they know what 'H' and 'O' stand for. *Being H_2O* is thus almost certainly part of the concept we associate with 'water.' In the version of Twin Earth that putatively establishes semantic externalism (the 1750 version) readers are asked to set aside the concept that they themselves associate with 'water' and imagine the concept that a pre-scientific speaker might associate with it (e.g., "the clear, drinkable liquid in the rivers and lakes etc."). We are then asked to make a judgment about what the extension of 'water' is in the example. What must be recognized is that performing this exercise requires great care. There is always a threat that we will sneak our own more familiar concept back into the example. In addition, we must not allow our knowledge that 'water' actually refers only to H_2O to influence our

---

which property regulates their use of 'right,' and further conversation will not move either from their normative theory, then "it seems increasingly reasonable to think that moralists and Twin-moralists would be warranted in interpreting each other as using different terms" (*ibid*. 228). Nevertheless, since both T^c and T^d are epistemically possible *for us*, it still (incorrectly) appears to us as if the Earthlings and their Twins are having a substantive disagreement.

> judgment about the semantic facts of the imagined Twin Earth case. If we do, then we are all but guaranteed to judge that the extension of t-'water' is different from the extension of 'water.' In such an event, our inability to set aside our own concept that we associate with 'water' would have given the false appearance of confirmation to semantic externalism.

This complaint shows that it is no virtue of PTE that it avoids the use of competing theories of water's composition. Indeed, this very feature threatens the integrity of the thought experiment. Our own judgment that XYZ is not water may be due to the fact that we cannot easily suspend our belief that water is in fact nothing but H_2O.

The defender of semantic externalism for natural kind terms could allay this worry by finding another Twin Earth story that involves competing theories of a (scientific) kind. This story must again yield the intuition that speakers on Twin Earth refer to a different kind (or express a different property) than Earthlings refer to using an orthographically identical term. Unfortunately for LM&D, if the externalist is successful in finding a replacement, then, in addition to having defended PTE, he will have shown that the use of competing theories of a kind does not give us reason to doubt our intuitions about Twin Earth cases. In that case, MTE is vindicated. On the other hand, if the externalist fails to find a suitable replacement for the twin water case, then this would suggest that Twin Earth thought experiments cannot be used to support semantic externalism. If so, PTE is a failure. But then, the observation that MTE lacks the very feature that renders PTE useless is hardly grounds for criticism against MTE.

2.4.5. Second reply.

Above, we saw that LM&D challenge the opponents of SEN to devise a compelling MTE story that does not make use of competing theories of *rightness*. I believe this challenge

can be met. Suppose that on Moral Triplet Earth 'morally right' is causally regulated by a property whose functional essence is specified by the following Nietzsche-inspired moral theory:[11]

> T^w: Necessarily, for any act, x, x is morally right iff by performing x, the agent of x expresses her will to power.

As in H&T's MTE story, we should suppose that Triplet Earthlings are normally disposed to perform actions they believe to have the "T^w property"[12] of *expressing will to power*. Moreover, they normally take an act's expressing will to power as an overriding reason for the agent to perform it. Let us add that Triplet Earthlings look upon those who fail to act in accordance with T^w with some sort of negative attitude (e.g., scorn or disgust). Agents who fail to act in accordance with T^w tend to feel some sort of negative attitude toward themselves (an attitude like shame, for instance). Importantly, such reactive attitudes do not generally accompany failures to act in altruistic ways; or, more precisely, these attitudes do not accompany failures to perform altruistic acts when those acts do not also express the agent's will to power.

I submit that Triplet Earthlings can use the predicate 'morally right' to engage in substantive moral disagreement with any Earthlings they might encounter. If we assume moral cognitivism, it follows that the Triplets' use of 'morally right' expresses the same property as is expressed by Earthling uses of 'morally right.' It does so despite the fact that the natural property that causally regulates the Triplets' use of 'morally right' is

---

[11] This example is inspired by David Copp's (1990: 247f). I do not mean to attribute T^w to Nietzsche himself.

[12] That is, the natural property whose functional essence is specified by T^W. Throughout this chapter, I also use 'T^c property' and 'T^d property' to denote the natural properties whose essences are specified by T^c and T^d respectively.

different from the one that regulates its Earthling uses. If this is correct, then CSN is again refuted.

Now, if Moral Triplet Earth is to satisfy LM&D's challenge, then it must also be true that (a) T^w is not an "epistemically live" theory of *moral rightness* and that (b) T^w supplies a potential criterion of *moral rightness* and not some other kind of *rightness* (e.g. prudential, or aesthetic). If (b) were false, then the defender of SEN could allow that Triplet Earthlings can have a substantive *practical* disagreement with Earthlings by using (tr-)'right' while denying that this disagreement is over which acts are *morally* right.

I am satisfied that both (a) and (b) are true. For me at any rate, T^w is not a live epistemic possibility. It is not among those theories in normative ethics that I can seriously entertain as true. I suspect this attitude toward T^w is widely shared by contemporary philosophers. On the other hand, the claim that (b) is true may face greater resistance. Some may take it to be a minimal requirement of a criterion of morally right action that the well-being of others is directly relevant to the *rightness* of any agent's act.[13] Call a criterion of right action that meets this requirement *other-regarding*. By T^w, the well-being of others is never directly relevant to the *rightness* of an agent's act. Hence, T^w is not other-regarding and, so the objection goes, it is not an eligible candidate for a criterion of *morally* right action. A further consequence is that Triplet Earthlings cannot use the term 'right' to have a substantive *moral* disagreement with Earthlings. Whatever else they are saying about an act when they apply 'morally right' to it, Triplet Earthlings are not ascribing *moral rightness* to it.

[13] Of course, the well-being of others may be *indirectly* relevant to the *moral rightness* of actions, given T^w. For example, there are likely to be situations in which an agent can express her will to power by forming alliances that benefit others.

Against an objection of this sort, I offer two observations. First, neither the *Oxford English Dictionary*, *The American Heritage Dictionary*, nor the *Merriam-Webster Dictionary* mention anything about other-regarding duties (or altruism) in their definitions of the adjective 'moral.' Second, it is common for moral philosophers to include ethical egoism on the menu of theories of morally right action (see, for example, Brandt 1959; Feldman 1978; Kagan 1998; Moore 1903; Ross 1930; Sidgwick 1907; cf. Foot 1958).[14] For an ethical egoist, the welfare of others is not directly relevant to the *rightness* of an action. Consequently, we would not expect egoism to be catalogued as a *moral* theory by these philosophers if *other-regardingness* were a requirement for any admissible theory or criterion of morally right action. I believe these two observations shift the burden of proof to those philosophers who would deny that T^w expresses a candidate criterion of *moral rightness*.

2.5. Isolating Moral Properties.

2.5.1. Second objection.

LM&D note that, given the fact that Moral Twin Earthlings are so similar to Earthlings, it is likely that the natural properties that regulate Earthling uses of 'good' and 'right' will be instantiated on Moral Twin Earth. Likewise, the natural properties that regulate t-'good' and t-'right' on Twin Earth will also be found on Earth (1999: 160). If I understand them correctly, the following may suffice to establish their point. According

[14] Philippa Foot's comments are especially relevant to the present discussion. In her "Moral Arguments" she considers various criteria that a proposition must meet in order to be counted as a moral proposition. She is emphatic that whatever criteria is adopted, it must count Nietzsche's doctrines as part of the subject matter of morality: "If a moral system such as Nietzsche's has been refused recognition as a moral system, then we have got the criteria wrong…We recognize Nietzsche as a moralist because he tries to justify an increase in suffering by connecting it with strength as opposed to weakness, and individuality as opposed to conformity" (1958: 33).

to the Earthling's moral theory, T^c, an act is right iff it has the property of *maximizing utility*. The same theory entails that an act is wrong iff it has the property of *not maximizing utility*.[15] Given bivalence, a T^c property (be it *maximizing utility* or *not maximizing utility*) will inevitably be instantiated by every action in the world. It makes no difference whether that action is performed in a social environment like Twin Earth where (it might be supposed) the moral agents do not take an active interest in the utility of their actions. Consequently, all of the Twin Earthlings' actions will inevitably instantiate a T^c property. Similar considerations apply, *mutatis mutandis*, for Earthlings and T^d properties. (I am assuming that, by T^d, an act is right iff it has the property of *treating no one as a mere means* and wrong iff it has the property of *treating someone as a mere means*). The upshot of all this is that any coherently described MTE case will have to include the presence of T^c properties and T^d properties on *both* planets. By contrast, according to the PTE story, Earth is entirely devoid of XYZ and Twin Earth is entirely devoid of H_2O. Here, then, we have another disanalogy between MTE and PTE.

According to LM&D, this feature of the MTE story raises trouble. If the T^c properties are ubiquitous on Twin Earth, and if the persons on Twin Earth really are psychologically similar to us, then surely they, like us, will take an interest in these properties just as we do. The problem is this:

> [I]f [Earthling] moral properties occur on Moral Twin Earth (and presumably play much the same roles that they play here), we should expect that the Moral Twin Earthlings have terms for them. The problem is that these sorts of considerations are likely to eclipse the facts in the Twin story about what properties "casually regulate" their use of terms like "good", "wrong", and so on. The business about

[15] In saying this, I am countenancing negative properties. Not everyone does so. I am uneasy about them myself. I help myself to negative properties here only because doing so makes it easier to see the basis for LM&D's claim that T^c and T^d properties exist on both planets. I believe their essential point could be made without appeal to negative properties. If not, and if negative properties really are indefensible, then so much the worse for LM&D.

> what causally regulates what is bound to be ignored [by readers], given the overwhelming likelihood that beings so similar to us would take an interest in [our] moral properties. Every other property they have lexicalized corresponds exactly to one we have lexicalized. Why stop short of [our] moral properties? (LM&D 1999: 160)

There are three key claims being made here: (i) T^c properties (such as *maximizing utility*) occurring on Twin Earth play "much the same roles" that they play on Earth, (ii) we should expect Twin Earthlings to have predicates that express T^c properties, and (iii) the expectation that Twin Earthlings possess such predicates distorts our understanding of the MTE scenario in such a way that we are misled into judging that Twin Earthlings really do express (e.g.) *maximizing utility* by t-'right' when, given the facts of the case, they do not. It is because of this confusion that we mistakenly judge that 'right' and t-'right' express the same semantic content and can be used to engage in substantive moral disagreement.

2.5.2. <u>Reply to (i).</u>

Claim (i) is unwarranted. If we stick to H&T's version of MTE, then we should not say that T^c properties play "much the same roles" on Twin Earth as they do on Earth. First, it is stipulated that T^c properties do not causally regulate Twin Earthlings uses of moral terms like 'right,' 'wrong,' and 'good.' So that is one respect in which the T^c properties do not play the same role on Twin Earth. More importantly, the role of T^c properties is rather different when it comes to the behavior of agents on Twin Earth. Given the MTE story that H&T tell, it is natural to suppose that, on Earth, an action's failure to maximize utility is taken by agents as strong grounds for avoiding it. We should also expect, furthermore, that when acts that fail to maximize utility are knowingly performed on

Earth, Earthling observers take a negative moral attitude toward the agent and the act. However, given H&T's stipulations, it seems clear that *not maximizing utility* does not play this role on Twin Earth. Twin Earthlings do not take an act's failure to maximize utility as strong grounds for avoiding it. Nor do they take a negative attitude towards acts and agents that fail to maximize utility. The property that plays that role on Twin Earth is the property of *treating someone as a mere means*. In short, T^c properties and T^d properties do not play "much the same roles" on both planets. Their roles are most saliently different with respect to the attitudes and behaviors of moral agents.[16] (Note the parallel with PTE: whereas on Earth the kind that plays the "watery" role is H_2O and not XYZ, on Twin Earth the kind that plays the watery role is XYZ and not H_2O.)

2.5.3. Reply to (ii).

Claim (ii) is also unwarranted. Indeed, I believe LM&D's assertion of (ii) is due to an oversight. To motivate it, they write that given

> …the assumption that Moral Twin Earthlings are like their Earthling counterparts in almost every respect…it's extremely natural to suppose that they have some way of referring to all the same sorts of things that we find significant, including [Earthling] moral properties. But if they have the ability to refer to these properties—properties that Earthlings take quite an interest in—there would have to be some special compelling reason to suppose that they did not refer to them (LM&D 1999: 60).

This claim overlooks the crucial detail that H&T include in the MTE story that prevents the Twin Earthlings from being *exactly* like their Earthling counterparts. In order to explain why different properties causally regulate Twin Earthling moral terms, H&T

[16] Of course, it may be that many of the Twin Earthling acts that instantiate T^d right-making properties also instantiate T^c right-making properties. But this should not trick us into taking T^c properties to play the same role on MTE that T^d properties play there. The difference in their roles is revealed when we consider how Earthlings and Twin Earthlings respectively would behave in either actual or counterfactual cases where an action instantiates a T^c right-making property but not a T^d right-making property (or vice versa).

stipulate that Earthlings and Twin Earthlings differ somewhat in their psychology. They offer only a hint as to what this difference might be:

> The differences in causal regulation, we may suppose, are due to species-wide differences in psychological temperament that distinguish Twin Earthlings from Earthlings. (For instance, perhaps Twin Earthlings tend to experience the sentiment of guilt more readily and more intensively, and tend to experience sympathy less readily and less intensively, than do Earthlings) (1990-91: 459).

This stipulation is included precisely to account for the fact that Twin Earthlings fail to take an interest in the properties that are so important to their Earthling counterparts. Although the stipulation is not developed in much depth, it seems plausible to suppose that some sort of psychological difference along these lines could result in Twin Earthlings taking an interest in different properties found in their natural and social environment than Earthlings take an interest in.[17]

Furthermore, a difference of interest along these lines could easily result in each planet's population failing to have a predicate that expresses the natural property that the other planet's population takes to be of moral importance. To take a less science fictional example, consider that a significant number of actual Earthlings take a very strong interest in the property of *being kosher*. Despite the lengths that some communities go to in order to consume only kosher animals, we would not be at all surprised to find a community of Earthling speakers who have no term that is translatable as 'kosher.' Nor would we find their lack of such a term more surprising upon our discovery that they have daily contact with kosher animals (e.g. bovines and chickens) and unkosher animals (e.g. pigs and shellfish). But if this is unsurprising in the case of *being kosher*, it is hard to see why it should be surprising in cases involving putative right-making natural

[17] For a review of psychological research that I believe lends empirical support to this speculation, see Richard Nisbett and Dov Cohen's *Culture of Honor* (1996).

properties. I conclude that if there is some reason why the stipulated difference in psychological temperament between Earthlings and Twin Earthlings is not sufficient to account for the former lacking predicates for T^d properties and the latter lacking predicates for T^c properties, then LM&D need to say what that reason is.

2.5.4. Reply to (iii).

But what if Twin Earthlings *did* have a predicate that expressed (a T^c property such as) *maximizing utility*? This would add yet another feature to the MTE scenario that is not shared by PTE. For, as LM&D note, in the PTE story it was supposed that XYZ entirely took the place of H_2O on Twin Earth; no H_2O was to be found there at all (1999:160). Because Putnam's Twin Earthlings had no H_2O in their environment, they presumably had no word to refer to it.

As we saw, LM&D contend that because it must be supposed in the moral case that Earthlings and Twin Earthlings have predicates that express each other's putative right-making properties, readers become confused in such a way that they (incorrectly) judge that t-'right' is used by Twin Earthlings to express the same content that Earthling uses of 'right' express. In particular, readers overlook all the details about the T^d property *treating no one as a mere means* causally regulating t-'right'. As a result, they erroneously suppose that t-'right' expresses *maximizing utility*, just as 'right' does.[18] The intended upshot here is that the judgment that 'right' and t-'right' share the same content should not be taken as a reflection of the semantic facts. It is instead a result of readers' inattentiveness to H&T's stipulations—where this inattentiveness is abetted by the fact

[18] Or perhaps LM&D meant to suggest that readers cannot help but to assume that there is some *other* natural property that both Earthlings and Twin-Earthlings express by 'right,' a property whose functional essence is captured neither by T^c nor T^d. This hypothesis is addressed Chapter 3.

that the MTE story tempts readers to suppose that Twin Earthlings possess predicates that express the same natural properties that causally regulate Earthling moral predicates.

Is there any reason to think that readers are likely to become so hopelessly confused by the supposition that Twin-Earthlings have predicates that express Earthling right-making properties (and vice versa)?[19] What is needed is another Twin Earth case against which we can calibrate our intuitions. Since it is taken as established that semantic externalism is true of chemical kind terms, I suggest that we devise another chemical kind Twin Earth story that embodies the allegedly suspicious features of the MTE story. If (contrary to what semantic externalism predicts) we find that we have the intuition that the target chemical kind term used by speakers on each planet has the same extension—and thus, their disagreements employing the term are substantive and not verbal—then we have reason to suppose that the MTE story is a flawed intuition pump. If, on the other hand, we have a firm intuition that the chemical kind term and its phonological twin have different extensions, then the differences between the MTE story and the PTE story that we are currently focused on give us no reason to doubt our intuitions generated by MTE.

Fortunately, we do not have to look far for a Twin Earth story with which we can calibrate our intuitions. Putnam had already provided one in "The Meaning of 'Meaning'." Just after finishing his discussion of the twin water case, Putnam asks his readers to consider another Twin Earth where the inhabitants use the substance molybdenum for all the purposes that we use aluminum for:

---

[19] A more charitable hypothesis about how readers are likely to deal with this supposition is that they will simply assume that if Twin Earthlings want to refer to Earthling T^c right-making properties, they (the Twin Earthlings) will simply use a non-moral, natural property name (e.g. 'maximizing utility') rather than a moral term such as 'rightness.' This hypothesis does no service for LM&D; and is not addressed in their paper.

> We will now suppose that molybdenum is as common on Twin Earth as aluminum is on Earth, and that aluminum is as rare on Twin Earth as molybdenum is on Earth. In particular, we shall assume that 'aluminum' pots and pans are made of molybdenum on Twin Earth (1975b: 225f).

In this Twin Earth story, Putnam stipulates that molybdenum is indiscernible from aluminum in its superficial properties. For our purposes, we may also add the stipulation that, here on Earth, the term 'aluminum' is causally regulated by the element with atomic number 13 (Al) and 'molybdenum' is causally regulated by the element with atomic number 42 (Mo). On Twin Earth, however, 'aluminum' is causally regulated by Mo, and 'molybdenum' is causally regulated by Al.[20] In line with this stipulation, we should suppose that the vast majority of Twin Earthling uses of 'aluminum' are applied to samples of Mo.

In the twin aluminum story, *both* Al and Mo are found on both Earth and Twin Earth. Moreover, speakers on each planet have terms that putatively refer to each of these kinds. Consequently, the twin aluminum story shares the feature of the MTE story that allegedly confuses readers. If MTE really is misleading as a result of this feature, then we should expect that our intuitions about the twin aluminum case will either run counter to our intuitions in the twin water case or else be held with less confidence. But this is not what we find. It is obvious that the extension of 'aluminum' when spoken by Twin Earthlings is different from its extension when spoken by Earthlings. Furthermore, when Earthlings say 'aluminum is composed of atomic number 13 atoms' and Twin

[20] For his own part, Putnam simply adds the stipulation that on Twin Earth 'aluminum' names molybdenum and 'molybdenum' names aluminum. At first sight, it would appear that this stipulation is question begging, since he uses this case to conclude that 'aluminum' and t-'aluminum' have different extensions. However, Putnam's interest in the aluminum example seems to be as a case where some members of each linguistic community know the underlying nature of the stuff they call 'aluminum' while others do not. If we needed to, the example could be modified so that no one on either planet knows enough chemistry to distinguish the two metals. The point presently being made would still stand.

Earthlings say 'aluminum is not composed of atomic number 13 atoms,' it is clear that their disagreement is merely verbal.

I conclude that LM&D have not shown that our intuitions are distorted by the feature that the twin aluminum story shares with MTE. If there is something misleading about MTE, it is not the fact that instances of T^c properties exist on Twin Earth. Nor is it the fact that Twin Earthlings are likely to have predicates to express such properties. It follows that there is no need for the opponent of SEN to redescribe the Moral Twin Earth story so as to purge it of these features.

2.6. Functional and Non-Functional Kinds.

2.6.1. Third objection.

LM&D write, "Another potentially distorting influence on the intuitions about Moral Twin Earth is the fact that moral properties are assumed to be functional properties. In contrast, the original Twin Earth thought experiment is framed in terms of non-functional natural kinds" (1999: 157). As I understand it, a property, P_1, is a functional property when an individual's instantiating P_1 depends upon that individual's instantiating another property, P_2, that "realizes" a certain causal role in that its environment. One of the interesting features of a functional property is that its instances may lack any intrinsic similarities with one another. What unifies these instances as instances of the same functional property is that they all share an extrinsic feature: they realize the same causal role. To take an extreme example, when instantiated in the right environments, *being a laser disk* and *being a filing cabinet* both realize a common functional property: *being an information storage device*. Nevertheless laser disks and filing cabinets share almost no

interesting intrinsic similarities. By contrast, I understand a non-functional kind (or property) to be a kind whose members (or instances) share a certain number of interesting intrinsic similarities (e.g., they share a similar physical structure and composition).

LM&D contend that our intuitions are less secure in Twin Earth cases involving a contrast between two (putatively different) functional properties than they are in cases involving non-functional natural kinds. The primary reason they offer for this insecurity is that, where functional properties are at issue, it may be unclear whether a functional predicate expresses a property, P_1, or some higher-level functional property, P_2, that is realized by P_1 in certain environments. The worry is that, because moral properties are assumed to be functional properties, readers may be led to suppose that 'right' and t-'right' both express a common higher-level property that is simply realized by different lower-level properties on each planet. That is, readers mistakenly suppose that there is a single functional property that is realized by *maximizing utility* on Earth and realized by *treating no one as a mere means* on Twin Earth. This possibility grounds LM&D's complaint that "H&T may gain some false leverage against ethical naturalism merely because at the crucial point in their argument, they compare ethical properties to non-functional natural kinds like water" (1999: 159).

2.6.2. Reply.

Let me begin by noting that defenders of SEN are in no position to criticize H&T for comparing moral properties to non-functional natural kinds like water. Many well-known defenders of SEN have themselves made this very comparison when defending their view (Boyd 1988: 196; Brink 1989: 157; 2001: 160; Lycan 1988: 200f; Railton

1989: 157f; Sturgeon 2003: 534). If there is a difficulty with extending to moral terms the semantics appropriate for non-functional natural kind terms like 'water,' then it is SEN's proponents (rather than its detractors) who need to resolve it.

More importantly, however, the PTE argument should be seen as *establishing*—rather than presupposing—that *being water* is a non-functional property. Putnam's readers have always had the option of judging that 'water' and t-'water' both express a single functional property found both on Earth and Twin Earth. On such a view, the essence of *being water* is given by a certain causal role. On Earth, the "watery role" is played by *being* H_2O. On Twin Earth, the same role is played by *being XYZ*. We can thus say that one and the same functional property of *being water* exists on both Earth and Twin Earth. The difference between the two planets concerns only the matter of which lower-level property realizes *being water*. On Earth *being water* is realized by *being* H_2O, while on Twin Earth *being water* is realized by *being XYZ*.[21] That this interpretation of the PTE story has always been available shows that the MTE story cannot be reproached for leaving a similar kind of interpretation available.

2.7. Conclusion.

LM&D along with Brink elucidate four apparent differences between the MTE thought experiment and the original PTE thought experiment. They claim that the respects in which MTE differs from PTE exert a distorting influence on our intuitions regarding the content of 'right' and t-'right.' I have argued that these differences are benign. If my

[21] Indeed, some philosophers actually have endorsed this reading of PTE, or something much like it. See, for example Zemach (1976) and Mellor (1977). Although they do not use the language of functional properties, both philosophers maintain that the extension of 'water' as used on both planets includes both H_2O and XYZ.

arguments are successful, then Brink and LM&D have failed to establish that MTE is any worse a thought experiment than PTE.

CHAPTER 3

MORAL TWIN EARTH AND HIGHER-LEVEL PROPERTIES

3.1. Introduction.

Towards the end of the previous chapter, we saw that LM&D raise the possibility that, within the MTE scenario, *moral rightness* is a single "higher-level" functional property that has distinct realizer properties on Earth and Twin Earth. They cite this possibility in order to cast doubt upon the clarity of our intuitions generated by MTE. However, following Eric Kraemer (1990-91), LM&D also entertain the hypothesis that *moral rightness* really is a higher-level functional property.[1] If true, this hypothesis appears to supply ethical naturalists with a different kind of answer to H&T's Moral Twin Earth argument against SEN.

In this chapter, I continue my defense of H&T's moral twin earth argument against SEN. I begin with a preliminary sketch of what I call the "higher-level properties reply" (HLPR) to MTE. Next, to get a clearer picture of the theoretical machinery upon which this reply depends, I outline a functionalist theory in the philosophy of mind. After that, I offer a more detailed statement of the HLPR to MTE. Finally, I argue that, while the metaethical theory that emerges from the HLPR avoids the chauvinistic conceptual relativism that results from the standard version of SEN, it is undone by a different form of relativism.[2]

---

[1] This reply to MTE is also suggested in Copp (2000), though Copp's ideas about how the reply should be spelled out appear to be different from those of Kraemer and LM&D.

[2] H&T (2000: 143) make the observation that the HLPR implies a kind of relativism in their reply to Copp's (2000).

3.2. A Prelimary Sketch of the Higher-Level Properties Reply to Moral Twin Earth.

Let us call the judgment that 'right' and t-'right' express the same content and can be used to engage in substantive moral disagreement 'the MTE intuition.' What explains why we have the MTE intuition? One possible explanation is that, as expressivists urge, we take the primary function of moral predicates to consist in expressing prescriptions, or speakers' attitudes, rather than intensions or properties of actions. According to this expressivist-type view, we should reject the cognitivist assumptions that we have granted the proponents of SEN. Earthlings and Twin Earthlings are able to have a substantive disagreement because both 'right' and t-'right' are used to prescribe or express approval of actions. An alternative explanation is that, as ethical non-naturalists urge, we take moral predicates to express non-natural properties that epiphenomenally supervene upon certain natural properties.[3] The non-naturalist can reject the claim that a predicate's being causally regulated by a natural property is sufficient (or even necessary) for that predicate to express that property.[4] She would then be free to assert that one (or even both) of the parties in the MTE scenario are simply mistaken about which actions have the non-natural property of *being right*.

---

[3] A non-natural property should be thought of as "epiphenomenal" in the sense that it is neither identical to, nor constituted by, the properties upon which it supervenes. Of course, this makes it difficult to explain why it is that non-natural properties co-vary in a law-like way with their subvenient natural properties. This difficulty gives rise to Mackie's charge that such properties are "queer" (1977: especially page 41). Arguably, non-natural properties are also epiphenomenal in the more traditional sense that they are causally inert (at any rate, their instances are causally inert). It is unclear to me, however, whether ethical non-naturalists would (or should) accept this latter characterization of moral properties.

It is worth adding that not all who call themselves non-naturalists will agree that moral properties are epiphenomenal. Shafer-Landau is one example. He agrees with synthetic ethical naturalists that moral properties are constituted by natural properties (2003: 72-78). Unlike ethical naturalists, however, he maintains that moral knowledge depends, in part, upon synthetic *a priori* intuition (*ibid*. Ch. 11).

[4] The non-naturalist would then owe us a different semantics for moral terms. I do not know what that semantics would look like. My brief comments about non-naturalism here are intended merely to acknowledge that MTE should not be seen as refuting all forms of moral realism.

There may be, however, another explanation of the MTE intuition that is available—an explanation that renders the MTE intuition compatible with CSN. In his commentary on H&T's (1990-91), Eric Kraemer recommends the following reply to MTE:

> One might claim that any moral theory that would suffice to capture the functional essence of a community which used the words 'good' and 'right' the way that earthlings do would have, whatever its ideological orientation (consequentialist or deontological), a functional core that would remain the same from population to population. Thus, the defender of CSN might claim that there is some ideology-neutral theory, T^n, which describes the functional essence of the core of any moral theory. The CSN supporter might suggest that T^c and T^d, though different in many obvious ways, both share T^n as a proper subset (1990-91: 469; cf. Laurence, Margolis and Dawson 1999:157ff).

Kraemer's proposal suggests what I call the "higher-level properties reply" (HLPR) to MTE. In its broadest sketch, the HLPR claims that a moral property like *moral rightness* is a multiply realizable functional property; in particular, it is a property whose essence is specified by an "ideology-neutral" theory).[5] On Earth, this functional property is realized by (but is not identical with) *maximizing utility*. On Twin Earth it is realized by (but is not identical with) *treating no one as a mere means*. Furthermore, we should understand that it is this "higher-level" functional property—and not the lower-level properties that realize it—that causally regulates the use of 'right' and t-'right.'[6] Thus, both of these

---

[5] As I suggest in §3.4 below, Kraemer's view seems to be that a theory is ideology neutral to the degree that its truth is compatible with the truth of a number of different competing first-order normative ethical theories.

[6] I should say something about my use of 'higher-level property' and contrast it with the notion of a second-order (or higher-order) property. I have noticed that some writers use these terms interchangeably. My use of 'higher-level property' corresponds to a picture of reality as divided into levels corresponding roughly to the various sciences. Some of these levels are "higher" in the sense that the items studied by the higher-level science supervene on the items studied by the lower-level science. From highest to lowest, it is common to rank the levels of reality as follows: psychology, biology, chemistry, and physics (this list is not meant to be exhaustive). With this picture in mind, we can view the higher-level property reply to MTE as claiming that moral properties are not to be identified with properties belonging to any of these levels. Instead, moral properties belong to a distinct level of reality that is higher than these others.

I understand a second-order property, roughly, to be a property that quantifies over other properties. A property is second order just in case it is the property of having a property of some sort. A

predicates express one and the same property. As a result, Earthlings and Twin Earthlings can use these predicates to engage in substantive moral disagreement with one another. In this way CSN is able to accommodate the MTE intuition.

To get a clearer picture of how this reply to MTE is supposed to work, it will be useful to reflect on functionalist accounts of mental properties.[7] In the next section, I take a brief detour through the philosophy of mind.

3.3. Functionalism about Mental Properties.

As Brink notes, functionalists about mental properties have typically thought that such properties are individuated by their causal roles:

> Mental states are identified and distinguished from other mental states in terms of the causal relations which they bear to sensory inputs, behavioural outputs, and other mental states. To take a hoary example, functionalist theories of mind claim that pain is identified and distinguished from other mental states by virtue of its tendency result from tissue damage, to produce an injury avoidance desire, and to issue in appropriate injury avoidance behaviour. The physical states which realise this functional state are the physical states upon which pain supervenes. (Brink 1984: 121; cf. Boyd 1980: 90; Jackson and Pettit 1988: 384; Shoemaker 1981: 263; Putnam 1967: 438).

Drawing on the example that Brink cites, let's sketch a toy theory of the mental property *being a pain*:

> T^p: A token state, x, is a pain iff there is a physical property, P, such that x has P and there is a tendency for token states that have P (i) to be caused by tissue damage in the organism in which the state occurs, and (ii) to result in aversive behavior in that organism.

---

second-order property need not be of a higher-level than the first-order properties it quantifies over. E.g., the *property of being a pain* may have the second-order property of *being disliked by most people*. Nevertheless, both properties belong to the level of psychology.

[7] Indeed, Kraemer cites functionalist theories of mental properties as inspiration for his reply to MTE (1990-91: 469).

To be sure, T^p is an impoverished theory of pain.[8] It is intended for illustrative purposes only.

Perhaps the most important feature of T^p for the present discussion is that it allows for pain to be multiply realized. Let me explain. Suppose that, in humans, brain states with the physical property of *being a C-fiber firing* tend to be caused by tissue damage and typically result in aversive behavior. Given T^p, we should say that C-fiber firings are pains. Next, let's suppose that there are creatures on Mars of a very complex sort. Following David Lewis's famous example,[9] let us suppose that, when a Martian undergoes tissue damage, this tends to cause the inflation of cavities in his feet. In turn, the inflation of the Martian's foot cavities tends to result in aversive behavior. For brevity, let's use 'being an FCI' to denote the property of *being a foot cavity inflation*. Given T^p, we should say that FCIs are pains. The property *being a pain*, then, is multiply realizable: in humans, it is realized by states with the property *being a C-fiber firing*; in Martians, it is realized by states with the property *being an FCI*. Despite their distinct realizing properties, Human pains and Martian pains are instances of one and the same higher-level property: *being a pain*.

Jackson and Pettit note that a functionalist theory like T^p is open to two readings (Jackson and Pettit 1988: 384f). Call the first reading of T^p, the "realizer reading." By the realizer reading, *being a pain* is identified with the property that plays the pain role—the "realizer" property. In our example, *being a C-fiber firing* and *being an FCI* are both

[8] The most obvious defect that comes to mind is the number of causal input and output conditions. Surely a plausible account of pain would include far more conditions. Furthermore, the most promising functionalist accounts of mental states include other mental states among the input and output conditions (see Lewis 1994). A final defect will be noted and corrected in §3.5. To anticipate: T^p needs to be modified to take account of the fact that certain lower-level properties may fail to realize *being in pain* if they are instantiated in the wrong system or environment.
[9] From his "Mad Pain and Martian Pain" (1980).

realizer properties; both play the pain role. Now, whatever the merits of the realizer reading, this is not how I understand T^p. In order to be coherent, the realizer reading of T^p requires acceptance of contingent identity (at least with respect to identities between mental and physical properties).[10] However, it is clear that the principal defenders of SEN reject contingent identity (Boyd 1980; Brink 1989: 157-159; Sturgeon 1985: 78n30). Moreover, given the semantics espoused by SEN's proponents, the realizer reading of T^p would lead us to deny that Martian FCIs are part of the extension of 'pain.' Here is why: If *being a pain* is identical to *being a C-fiber firing*, then presumably our use of 'pain' is causally regulated by the property of *being a C-fiber firing*. But if that is true, then, given Boyd's semantics, something is in the extension of 'pain' at a world just in case it has the property of *being a C-fiber firing*. Since FCIs necessarily lack the property of *being a C-fiber firing*, we should conclude that FCIs are in neither the extension nor intension of 'pain.' But this is wrong. If we see a Martian writhing on the floor after undergoing tissue damage and experiencing an FCI, we speak truly when we utter 'The Martian is in (or has a) pain.' This could not be so if FCIs were excluded from the intension of 'pain,' as would be the case given the conjunction of the realizer reading of T^p and Boyd's causal-regulation semantics.

For present purposes, then, we should read T^p in the alternative way outlined by Jackson and Pettit. By the alternative "role reading" of T^p, *being a pain* is identified with the second-order property of *having a property that plays the pain role*. In other words,

[10] Here's why. Assume for *reductio* that identities between mental and natural properties are not contingent. Such identities, then, are necessarily true, if true at all. Now assume that *being a pain = being a C-fiber firing*. By the necessity of identity, it follows that in *all* possible worlds, *being a pain = being a C-fiber firing*. However, we have also been supposing that pain is multiply realizable. Given the realizer reading of T^p, this means that there is a possible world, w, in which, *being a pain = being an FCI*. By the transitivity of identity, we are forced to conclude that, at w, being an *FCI = being a C-fiber firing*. But this is absurd. We have been supposing that these are distinct properties. Thus, if we were to suppose that pain is identical to its realizer properties, we should take these identities to be contingent.

something has the property of *being a pain* just in case it has some *other* property (e.g., being a *C-fiber firing*, *being an FCI,* etc.) that plays or realizes the pain role. If we allow that this second-order property fixes the intension of 'pain,' then Martian FCIs fall within that intension, as they should. For such states instantiate the property of *having a property that plays the pain role*, just as human C-fiber firings do. (Importantly, on the role reading, it is a mistake to identify *being a pain* with either of its realizer properties. It is a distinct property.)

3.4. The Higher-Level Properties Reply to MTE.

Although Brink has expressed a willingness to view moral properties as multiply realizable functional properties (1984: 121; 1989: 157-159), it should be recalled from §1.6.1 that he and his fellow SEN proponents suppose that the essence of *moral rightness* will be specified by a theory in first-order normative ethics. It is this assumption that, when conjoined with CSN, leads to the conclusion that 'right' and t-'right' express different properties. The central insight of the HLPR is that this assumption must be abandoned. Instead, the defender of SEN should suppose that the essence (and functional role) of *rightness* is to be specified by a theory of a more "metaethical" flavor—a theory that is to a large extent neutral between first-order normative theories (this is what I take Kraemer to mean when he says the functionalist account of *rightness* should have an "ideology-neutral" core). In §3.7 I offer a sketch of the sort of neutral theory that I believe the HLPR requires. For the moment, I would cite ideal observer theories as familiar examples of metaethical accounts of *rightness* that are, in important respects,

neutral with respect to first-order normative theories.[11] Whatever theory is adopted, it must be such that it is possible for the *rightness* role to be played (or realized) by *maximizing utility* on Earth and played by *treating no one as mere means* on Twin Earth. *Rightness*, however, is not to be identified with either of these realizer properties that play the *rightness* role.[12] Like *being in pain*, *rightness* is to be thought of as the second-order property (that is of a higher-level than its realizers); it should be thought of as something like the property of *having a property that plays the rightness role*.

The HLPR to MTE can now be stated. In their thought experiment, H&T stipulate that the essences of the properties regulating the use of moral terms will be specified by theories in first-order normative ethics. Given the revision that Kraemer proposes for SEN, H&T are no longer entitled to this stipulation: by the new version of SEN, we expect that a higher-level, ideology-neutral property will causally regulate the use of moral terms. With this new understanding of what sort of property causally regulates the use of moral terms, the MTE intuition—the intuition that 'right' and t-

---

[11] This requires some comment. First, it is controversial whether ideal observer accounts of moral properties are realist accounts. The trouble is that they make moral facts stance-dependent: the standard that fixes the moral facts is made true by the beliefs and attitudes of the ideal observer. If so, then it would seem that proponents of SEN (who are realists) cannot avail themselves of an ideal observer theory of *moral rightness* in pursuit of a HLPR to MTE. Second, there may be some controversy as to whether ideal observer theories really are (in the required respect) neutral with respect to first-order normative theories. Indeed, it might be argued that an ideal observer theory is itself a competing theory in first-order normative ethics. After all, like theories in normative ethics, ideal observer theories supply us with a criterion for morally right action (e.g.: "An act is morally right iff it would be approved of by an ideal observer"). What is important for our purposes, however, is that a theory of this sort need not be seen as a competitor to utilitarian and deontological accounts of right action. For it could turn out, upon investigation, that ideal observers approve of all and only those acts that maximize utility. In that case, utilitarianism would be true alongside the ideal observer theory. Similarly, it could turn out that ideal observers approve of all and only those acts that treat no one as a mere means. In this way, the ideal observer theory is at least *prima facie* neutral with respect to (at least some of) the competitors in normative ethics. For illustrations of ideal observer theories of *rightness* see Firth (1952) and Smith (1994).

[12] Nor should *rightness* be identified with any other natural property that is treated as right-making by theories in first-order normative ethics. Note that, in this respect, the functionalist account of *rightness* being developed here differs from the moral functionalism of Frank Jackson (1998). Jackson identifies *rightness* with its realizer properties. He suggests, for example, that *rightness* might turn out to be identical to *maximizing expected hedonic value* (*ibid*. 141-143).

'right' have the same content and can be used to engage in substantive moral disagreement—is rendered compatible with CSN. For all H&T have said (minus the contested stipulation) it may be that there is a single higher-level property that causally regulates the use of both 'right' and t-'right.' If so, then, by CSN, these predicates share the same semantic content.[13] The remaining differences between the Earthlings and the Twin Earthlings can be accounted for by the hypothesis that this higher-level property is realized by one natural property on Earth and a different natural property on Twin Earth. In this way, defenders of SEN can accommodate the MTE intuition without abandoning their preferred externalist moral semantics embodied in CSN.[14]

In the next section, I consider some difficulties facing the synthetic ethical naturalist who adopts the HLPR.

---

[13] Notice that the HLPR cannot work unless it is denied that that the lower-level realizer properties are what causally regulate the uses of 'right' and t-'right.' This is worth taking note of, since one might be tempted to state the HLPR this way: although *distinct* natural properties causally regulate the use of 'right' and t-'right,' they nevertheless share the same semantic content since the both of properties that regulate these predicates realize the same, single higher-level property. The problem with this way of putting the HLPR is that the mere fact that a lower-level property that causally regulates the use of a predicate, 'F,' happens to realize a higher-level property does not, as I understand CSN, make it true 'F' expresses the higher-level property rather than the lower-level property. To see why not, consider the predicate 'C-fiber firing.' That *being a C-fiber firing* realizes the higher-level property of *being a pain* does not make it the case that 'C-fiber firing' expresses the property of *being a pain*. For if that were the case, we would speak truly when we say of a Martian writhing on the floor after undergoing tissue damage, "That Martian is having (or experiencing) a C-fiber firing." But this is wrong. For all that has been said, the Martian's body may not even contain any C-fibers.

[14] Of course, the opponent of SEN might return with a revised MTE case. Here, we might imagine that two different ideology-neutral, higher-level properties regulate the use of 'right' and t-'right.' I will leave this avenue unexplored for three reasons. First, it isn't obvious to me just how to develop such a case. Second, even if I did see how to develop it, I believe it would require more space than I can afford here. Third, I believe that, even if such a case could be spelled out, the judgment that 'right' and t-'right' express the same property—despite being causally regulated by distinct ideology-neutral, higher-level properties—would not be all that compelling.

3.5. Troubles for the Higher-Level Property Reply: Agent's-Group Moral Relativism.

At first glance, the HLPR looks to be a neat solution to the challenge of MTE. Unfortunately, it comes at a high price for moral realists. Moral realism is typically intended as a form of moral absolutism. Indeed, Brink himself characterizes moral realism as an "antirelativist metaethical view" (1989: 26). In light of this, it should come as an unpleasant surprise—though perhaps it will be obvious to the reader by now—that the HLPR in defense of CSN carries a commitment to moral relativism.[15] Let me explain.

In my example involving mental properties, I suppressed one important detail. We should not say, without qualification, that *being a C-fiber firing* realizes *being a pain*; for there may well be C-fiber firings that do not realize pains. This could occur if there are creatures in which C-fiber firings play an entirely different neurological role than they play in humans. Suppose that, in octopi, instances of *being a C-fiber firing* are never caused by tissue damage and never result in aversive behavior. Instead the C-fiber firings in octopi tend to be caused by sensing prey and tend to result in hunting behavior. It is obvious that, in the octopus, *being a C-fiber firing* does not realize *being a pain*.[16] The lesson here is that a property realizes the functional role of pain only with respect to certain "systems" or environments. In this example, the relevant environment would be a

[15] In fairness, Kraemer appears to recognize that some form of relativism will result from his proposal, though his acknowledgement is both indirect and buried in a footnote in the conclusion of his commentary (safely away from the section of text where the higher-level property proposal is actually presented). There, he calls David Lewis's functionalism about mental properties a "relativistic approach." He does not, however, explicitly acknowledge that the relativism might carry over to the functionalist account of moral properties (1990-91: 472n10).

[16] It is obvious, at least, given T^p. But note that T^p makes no claim concerning the qualitative feel of pain states. Philosophers of mind who emphasize the role of qualitative feel in individuating mental properties may have lingering doubts that the octopus's C-fiber firings aren't pains. For them, let me add the stipulation that octopus C-fiber firings do not have a similar qualitative feel to human C-fiber firings.

particular brain—or, perhaps more narrowly, a particular region of a particular brain. To make this explicit, we should revise our toy theory of pain as follows:

> T^{p2}: A token state, x, is a pain iff there is a physical property, P, such that x has P in a system, S,[17] and there is a tendency for token states that have P in S (i) to be caused by tissue damage in the organism in which the state occurs and (ii) to result in aversive behavior in that organism.

Similar considerations should extend to the functionalist account of *rightness* under discussion here. It is not sufficient to say that an act is right just in case it has whatever property plays the *rightness* role. Given the MTE scenario, more than one property plays the *rightness* role. What a functionalist account of *rightness* should say is that an act is right just in case it has whatever property plays the *rightness* role in the relevant system or environment in which the act is performed. Presumably, the systems in question will be social groups or social environments. Thus, it should be denied that the property of (e.g.) *maximizing utility* realizes *rightness* simpliciter. Rather, it realizes *rightness* only in those social environments where *maximizing utility* plays the *rightness* role. In the MTE story, it is plausible to suppose that *maximizing utility* plays the *rightness* role on

---

[17] By 'system' I mean 'system-token.' In this case, a system-token would be a particular individual's brain at a time. It should be noted that, by relativizing instances of pain to system (i.e., brain) *tokens*, T^{p2} is unable to accommodate the phenomenon that Lewis calls "mad pain." The madman is a human being whose C-fiber firings have different causes and effects than that of nearly all other humans. For the madman, C-fiber firings are caused by "moderate exercise on an empty stomach" rather than tissue damage. Moreover, his C-fiber firings result in him concentrating on mathematics, rather than engaging in aversive behavior. In Lewis's view, the madman's C-fiber firings are pains. This is so despite the fact that the madman's C-fiber firings do not play the same sort of causal role that they play in his fellow human beings. To accommodate Lewis's judgment that the madman is in pain when he undergoes C-fiber firings, T^{p2} would need to be reformulated. In particular, where T^{p2} quantifies over system-tokens, the reformulated theory should quantify over system-kinds. Something along the following lines might do the trick:

> T^{p3}: A token state, x, is a pain iff there is a property, P, such that x has P in a system of kind, K, and there is a tendency for token-physical states that have P in systems of kind K (i) to be caused by tissue damage in the organism in whom the physical state is realized and (ii) to result in aversive behavior in that organism.

The madman's brain, for all its odd wiring, is still a human brain. Since C-fiber firings realize pain in human brains, the madman's C-fiber firings constitute pains according to T^{p3}. (For an argument against mad pain, see Shoemaker [1981: 267-272].)

Earth but not on Twin Earth. Likewise, it is plausible to suppose that *treating no one as a mere means* plays the *rightness* role on Twin Earth but not on Earth.

If the functionalist theory of *rightness* is understood in this fashion, then it carries a commitment to "agent's-group moral relativism." David Lyons describes agent's-group moral relativism (hereafter, "agent-relativism") as the view that "an act is right if, and only if, it is permitted by the norms of the agent's group" (1976: 109; cf. Sturgeon 1994: 83).[18] It should be clear that, given the functionalist account of *rightness*, the answer to the question of which norms apply to a given agent depends upon which property realizes the *rightness* role in that agent's social environment. Because *maximizing utility* realizes *rightness* in the Earthlings' environment, the moral status of Earthlings' actions depends upon consequentialist norms; and because *treating no one as a mere means* realizes *rightness* in the Twin Earthlings' social environment, the moral status of Twin Earthlings' actions depends upon deontological norms.

To illustrate the sort of moral relativism at issue, consider the performance of an organ harvest on Earth. (Recall that the organ harvest is an act that maximizes utility by treating someone as a mere means.) Since *maximizing utility* realizes *rightness* on Earth, the Earthling organ harvest is morally right. Suppose, however, that a duplicate organ harvest is performed on Twin Earth.[19] There, *rightness* is realized by *treating no one as a mere means*. Since the Twin Earthling organ harvest lacks this latter property, it is not morally right. It is immaterial that the act also maximizes utility: *maximizing utility* does

[18] Some may find it desirable to amend Lyon's formulation of agent-relativism with the further claim that, as a matter of contingent fact, different social groups have different norms. As it stands, his formulation is consistent with there being a single set of norms that applies to all moral agents. It is unclear to me whether we should describe such a scenario as one in which moral relativism prevails.

[19] And, moreover, suppose that this duplicate act also maximizes utility. (This stipulation needs to be added since *maximizing utility* is an extrinsic property of actions and need not be shared among duplicates.)

not realize *rightness* on Twin Earth. It would seem, then, that a consequence of the higher-level property reply to MTE is that two actions alike in all salient non-moral features may nevertheless diverge in their moral status. The moral status of either act depends largely upon features of the social environment in which it was performed. In particular, it depends upon those features of the social environment that determine which natural property plays the *rightness* role there.[20]

3.6. "Merely Possible" Relativism.

Although the proponent of the HLPR must allow that a natural property realizes *rightness* only relative to a particular social environment, he may deny that this entails a full form of agent-relativism. As we saw, the relativization clause in the formulation of the functionalist theory of *rightness* is needed in order to deal with possible worlds like the

---

[20] Could it simply be denied that the functionalist account of *rightness* needs to include the sort of relativization to environments that is needed by the functionalist account of *being a pain*? I doubt it. Such a denial would commit us to the claim that a natural property that realizes *rightness* in one social environment realizes *rightness* in all social environments. Thus, an act-token is right iff it has a property (any property) that plays the *rightness* role anywhere in the world. In the confines of the MTE story, this entails that, on either planet, an act is right iff performing it *either* (a) maximizes utility or (b) treats no one as a mere means. Because the same disjunctive standard of *rightness* applies to all moral agents in all environments, this understanding the functionalist account of *rightness* avoids agent-relativism.

For those who do not immediately find this strategy implausible, I offer the following as reason to reject it. Hitherto, I have mentioned only the predicate 'morally right' in my examples. However, a full functionalist account of moral properties will need to say something about the properties expressed by 'morally wrong' and 'morally obligatory.' I take it that, for the functionalist, an act is wrong (obligatory) iff it has a property that plays the *wrongness* (*obligation*) role. Given the utilitarian morality of Earth, we should suppose that *wrongness* is realized by the property of *failing to maximize utility*. On the other hand, given the deontological morality of Twin Earth, we should also suppose that wrongness is realized by the property of *treating someone as a mere means*. By the strategy under consideration, it seems that we should say that, in the confines of the MTE story, an act is wrong iff performing it either (a) fails to maximize utility or (b) treats someone as a mere means. The trouble is, with these disjunctive criteria of *rightness* and *wrongness*, it is possible for one and the same act to be both right and wrong at once. For example, an act of performing an organ harvest is right because it maximizes utility; but it is also wrong because it treats someone as a mere means. We are now in the throes of a practical contradiction. To make matters worse, when combined with an independently plausible deontic principle, the practical contradiction becomes a logical contradiction. The deontic principle in question is simply that, if an act is wrong, then it is not right to perform it. From this, it follows that it is right to perform the organ harvest and it is not right to perform the organ harvest. But that is absurd. The present proposal for avoiding agent-relativism, then, is incoherent.

one containing the MTE-scenario. Still, the proponent of the HLPR may insist that, as a matter of contingent fact, here on Earth in the actual world (where Twin Earth does not exist) there is one and only one property that realizes *rightness*. Thus, as a matter of contingent fact, there is a single standard of *moral rightness* that applies to all actual Earthlings. It might be claimed that this is all the absolutivity that a moral realist need ask for. That there could have been multiple moral standards—or even that there may actually be other standards on distant planets we will never visit—is no cause for alarm. The only relativism this realist is bothered by is the intra-planetary and intra-worldly variety.

It is worth observing that there is really no need for H&T to stipulate that Earth and Twin Earth are separate planets. That stipulation, I believe, is a mere tip of the hat to Putnam's original Twin Earth thought experiment. The MTE story could be easily be recast using socially isolated groups of Earthlings.[21] Consequently, the realist must assure us that such a scenario is not how things actually are. That is to say, he must assure us that, although *rightness* is capable of being realized by different natural properties in different social groups on Earth, as a matter of fact, among all actual Earthling social groups, there is only one *rightness* realizer. Whether or not this assurance is to be believed will depend upon what the best functionalist account of *rightness* turns out to be and whether the empirical evidence shows the *rightness* role to be realized by a single property for all Earthling groups.

Given our current evidence, I see little reason for optimism on the part of the realist. The existence of deep and seemingly irresolvable intercultural disagreement

[21] R. M. Hare, for example, offers an Earthbound MTE-type story involving a missionary and cannibals (Hare 1952: 148ff); and, of course, my example in Chapter 1 (§1.5.4) involving the Benthamanians and New Immanuelers provides the template for another story of this kind.

about which acts are morally right has long been cited as evidence that there is no single standard of *rightness* that applies to all humans (Mackie 1977: 36ff; Westermarck 1932/1960: Ch. 7). While Brink and others have offered alternative explanations of moral disagreement that are more acceptable to moral absolutists (Brink 1989: 197-210; Boyd 1988: 212-214), it cannot be denied that the relativist's preferred explanations are dramatically favored once the concession is made that *rightness* is multiply realizable in the way that the HLPR requires. Surely, given this concession, the most charitable explanation for why different cultures behave as if *rightness* is realized by a different natural property than the one that we believe realizes it is that, for some of those cultures anyway, *rightness actually is* realized by a different property. The alternative explanation—the one needed by the realist—is far less charitable. It requires us to hold that all of those cultures who disagree with us are simply mistaken about the moral facts or relevant non-moral facts. Of course, charity is not the only theoretical virtue to be considered when weighing competing explanations of persistent moral disagreement. Still, it seems to me that once we accept the view that *rightness* is multiply realizable, the difference in charity between the competing explanations is so dramatic that it shifts the presumption in favor of the relativist's explanation (especially for those cases where the moral disagreement does not obviously result from one party being mistaken about the non-moral facts).

3.7. Is Agent-Relativism Compatible With Moral Realism?

It seems likely, then, that a proponent of SEN who adopts the HLPR will find himself committed to (actual) agent-relativism. At any rate, he is certainly committed to agent-

relativism with respect to hypothetical cases such as the MTE scenario. Although Brink has characterized moral realism as an anti-relativist view, nothing that has been said so far shows that realism logically precludes relativism.[22] If it does not, then there may be moral realists who find agent-relativism to be a tolerable price to pay for the HLPR. In the present section, I argue that there is good reason to believe that the functionalist account of *rightness* needed for the HLPR is incompatible with moral realism.[23]

The reason moral realism is thought to preclude moral relativism is that standard forms of relativism violate realism's stance-independence clause. In §1.2.1 we saw that in order to satisfy the stance-independence condition, it must be the case that the moral standard that fixes the moral facts is not made true in virtue of its being "ratified" by some actual or hypothetical appraisers. As Sturgeon's characterizes it,

> …we ought not to count a view as realist unless it holds that these moral truths are in some interesting sense *independent* of the subjective indicators—our moral beliefs and moral feelings, as well as moral conventions constituted by coordinated individual intentions—that we take as guides to them (1986b: 117; cf. Boyd 1988: 182; Brink 2001: 154).

It is important to note that the kind of ratification that is relevant here need not be thought of as a conscious activity of appraisers; their mere tacit acceptance of the relevant standard suffices to render moral facts dependent on "subjective-indicators" of the sort that Sturgeon mentions.

For illustration, consider a version of moral relativism that holds that the moral status of an act depends upon whether or not it is permitted by a code of rules that are

---

[22] For arguments to the effect that moral realism can be reconciled with relativism, see Oddie (1999) and Sayre-McCord (1991)

[23] As far as I can see, the argument of the present section does not depend upon the success of my rebuttal to the objection raised in §3.6. Here, I argue that whatever functionalist account of *rightness* the realist adopts for the HLPR, it will violate the stance independence clause. Such a violation could occur even if, as a matter of contingent fact, only one property actually realizes *rightness*.

accepted or endorsed by the members of a given social group. That a view like this violates the stance-independence clause can be seen when we consider that the moral status of the act would have been different had the members of the group accepted or endorse a different code of rules. The truth of this counterfactual proposition is evidence that moral facts, as conceived by this version of relativism, are not independent of subjective indicators, as the stance-independence requirement demands.

The question before us, then, is whether the functionalist account of *rightness* needed for HLPR can preserve the stance-independence of facts about which actions are right. I believe that it cannot. When the functionalist account of *rightness* is fully spelled out, it will be seen to entail that moral facts depend upon subjective indicators. If so, the HLPR is incompatible with moral realism and, hence, cannot be adopted as a means of defending the SEN brand of realism from the MTE argument. While I cannot attempt to offer a full functionalist account of *rightness* here, I hope to say enough to show why it is doubtful that moral facts would be stance-independent on the sort of account needed for the HLPR.

Recall from §3.3, that Brink understands a functional property to be individuated by its causal role. A functional account of *rightness*, then, will involve a specification of the causal role that this property plays. As with our toy account of pain, the causal role of *rightness* will be articulated by a theory specifying the causal relations that right acts stand in with respect to certain inputs and outputs. Assuming the sort of realism about moral properties that SEN proponents favor, let us ask what sorts of things cause and are caused by morally right actions.[24] The most obvious answer, it seems to me, is that moral

[24] I expect that many, including some moral realists, will be skeptical of the claim that moral properties exert a causal influence on anything (e.g. Nagel 1986: 144; Shafer-Landau 2006: 225). Nevertheless all of

facts cause and are caused by human behaviors, thoughts, and attitudes.[25] After all, right actions are *actions*; and the standard causal explanations of actions invoke the beliefs, desires and intentions of those creatures that perform them (Davidson 1963). It stands to reason, then, that the beliefs, desires and intentions that drive human behavior will figure in the causal profile of morally right actions. In light of this, we might offer as a causal input clause the claim that sympathetic or impartial agents tend to prefer the performance morally right acts to their alternatives.[26]

For a causal output clause, we will need to cite the sorts of events that result from morally right actions. Among the most obvious causal consequences of an action's being wrong is that, often, impartial observers take a negative attitude toward both the action's performance and the agent who performed it. Indeed, the agent herself often takes a negative moral attitude towards her act (e.g., she feels guilt). In turn these negative attitudes tend to give rise to behaviors such as the condemning of the act, or the punishing of the agent. This suggests as an output clause the claim that morally wrong acts tend to cause observers—at least, those observers who are sympathetic and impartial—to condemn and take a negative moral attitude toward the act and its agent.

---

the principal synthetic ethical naturalists affirm the causal efficacy of moral properties. Indeed, if moral properties do not enter into causal relations, then we have a quick refutation of CSN. For in that case, moral properties could not be what causally regulate our use of moral terms. Moreover, in light of their commitment to EC, it is hard to see how they could countenance unreduced moral properties unless those properties had some sort of causal profile.

A more general worry that some may have is that it simply makes no sense to speak of any property—moral or non-moral—as having causal powers. Behind such a concern is the thought that properties are abstract objects and abstract objects cannot enter into causal relations. One way to deal with this worry is to translate talk about the causal powers of a given property, P, into talk about the causal powers that concrete individuals have in virtue of instantiating P.

[25] Below, I will consider human welfare as another item that stands in a causal relation to right acts.

[26] Although he takes it to be an output clause, Frank Jackson acknowledges something like this principle in his own sketch of a functionalist account of *rightness*: "The judgment that an act is right is normally accompanied by at least some desire to perform the act in question…" (1998: 131; Cf. Smith 1994: 39).

Using these proposed input and output clauses, let us formulate a toy functionalist theory of *moral rightness*.

> T^n: An act-token x is morally right iff there is a natural property, P, such that x has P in system S and there is a tendency for (i) act-tokens that have P in S to be preferred to their alternatives by sympathetic and impartial agents, and (ii) act-tokens that lack P to elicit condemnation and negative moral attitudes on the part of impartial observers.

As with the toy account of pain, we are to identify *rightness* with the property of *having a natural property that plays the rightness role*; *rightness* is not to be identified with the realizer properties that are quantified over by 'there is a natural property P' (e.g. natural properties like *maximizing utility* and *treating no one as a mere means*).

If T^n (or something relevantly like it) were our best higher-level functionalist account of *rightness*, then the HLPR implies the falsity of moral realism. The reason for this is that, by T^n, the standard that fixes the moral facts is itself determined by subjective indicators (i.e., moral attitudes and beliefs). For example, given T^n, it is plausible to suppose that the reason *rightness* is realized by a different property on Twin Earth is that, unlike Earthlings, Twin Earthlings tend to prefer actions that treat no one as a mere means and take a negative moral attitude to actions that do treat someone as mere means. Furthermore, given T^n, if we Earthlings had sufficiently different preferences and moral attitudes, a different natural property would make our actions right than the one that currently makes them right. By rendering moral facts dependent upon our attitudes and preferences in this way, T^n violates the stance-independence requirement for moral realism. It would appear, then, that moral realists cannot avail themselves of the HLPR to MTE.[27]

[27] Note that functionalist accounts of *rightness* that treat the lower-level realizer property as identical with *rightness* do not have this problem and are compatible with moral realism. For whether or not (e.g.) an act

To save the HLPR, the realist needs a functionalist account of *rightness* that does not determine which natural properties realize *rightness* by appeal to the attitudes that agents and observers take towards actions with those properties. Conspicuously absent from my list of things that cause and are caused by right actions is any mention of the causal impact right action has on human welfare. Perhaps we should look here for a realist-friendly functional account of *rightness*. Consider Brink's own suggestion for the functional role that *moral goodness* plays:

> [T]he realist might claim that moral properties are those which bear upon the maintenance and flourishing of human organisms. Maintenance and flourishing presumably consist in necessary conditions for survival, other needs associated with basic well-being, wants of various sorts and distinctively human capacities (1984: 122).

One might turn this into an output clause for an account of *rightness* by supposing that when an act is right, it has a natural property, P, and occurs in a social environment, S, such that there is a tendency for acts with P in S to promote the flourishing of human organisms. Call this *the flourishing condition*.

One worry about the flourishing condition is that, on some ways of understanding it, it threatens to commit us to consequentialism of some form.[28] However, in order for the functionalist account to serve the HLPR, it must be such that a deontological property (such as *treating no one as a mere means*) can realize the *rightness* role. Otherwise, t-'right' would not count as being causally regulated by the same higher-level property as 'right.' In that case, the reply to MTE collapses. The same difficulty will arise for any

---

maximizes utility does not depend upon the attitudes of observers in a way that violates stance-independence. Thus, if *rightness* is identified with maximizing utility, *rightness* itself (or facts about which acts are right) counts as stance-independent.

[28] Indeed, it looks as though it could force us to accept a normative view according to which an act is morally right just in case it is permitted by the moral code whose currency in our social environment would maximize human flourishing.

other functionalist account that the realist offers insofar as that account fixes right-making properties by input and output clauses that embody substantive moral principles.

Of course, there is no harm in including *some* substantive principles among the inputs and outputs in our functionalist account, as long as those principles are construed in a broad enough way that the *rightness* role could be realized by any number of potential right-making properties. Surely, even a deontologist could admit that right actions *tend* to promote human flourishing (even if not all right acts do so). Unfortunately, this means that a property's satisfaction of the substantive input and output clauses will be (indeed, must be) insufficient to fix a determinate natural property as the *rightness* realizer in a given environment. To yield determinate moral facts, the functionalist account of *rightness* must be supplemented with ideology-neutral input and output clauses. It seems to me that such clauses will need to make reference to the attitudes, beliefs, or behavior of agents within a social environment. I cannot see any other options. If I am right, then even a functionalist account of *rightness* that includes substantive input and output clauses will violate realism's stance-independence requirement; it remains true on this sort of account that, if our attitudes (or other subjective indicators) had been different, a different natural property would have realized *rightness*.[29]

[29] Horgan and Timmons themselves recognize in their (2000) that naturalist moral realists confront a dilemma of the sort described in this section. They write, "The first horn is that the putatively reference-fixing relation *R* might fail to fix *determinate* reference-relations between moral terms and certain natural properties because there are too many eligible natural properties that satisfy the constraints imposed by *R*. […] The second horn of the dilemma arises if one grants that the proposed reference-fixing relation *R* suffices to pin down some unique class of natural properties as the putative referents of moral terms." In that case, the realist's semantics falls to an MTE counterexample (H&T 2000: 240; cf. 1996a: 32-34).

It appears, then, that the realist cannot avail himself of the HLPR in order to defend CSN since the HLPR requires a functionalist account of *rightness* that is incompatible with realism.

3.8. Conclusion.

The HLPR is a natural and tempting answer to the challenge that MTE poses to Boyd's causal semantics for moral terms (CSN). In this chapter, I hope to have given a clear picture of what the HLPR is and what its commitments are. In particular, I have argued that the adoption of the HLPR carries a commitment to agent-relativism. In addition, I have argued that the HLPR requires a functionalist account of moral properties that is incompatible with moral realism. Since CSN is of interest to us primarily for its role in the defense of the naturalist moral realism, the HLPR turns out to be self-defeating: it preserves CSN at the cost of abandoning moral realism.

CHAPTER 4

BRINK'S MORAL SEMANTICS

4.1. Introduction.

Until now I have been supposing that the theory of content-fixing that underwrites SEN's externalist semantics is Boyd's causal regulation account. As we have seen, this account is vulnerable to H&T's Moral Twin Earth argument. H&T contend that the MTE thought experiment can be used to undermine any externalist moral semantics insofar as that semantics succeeds in pinning down a determinate (and realist-friendly) semantic content for moral predicates. They claim that any such theory will lead either to a form of chauvinistic conceptual relativism, or else to a form of agent (or "standard") moral relativism that is in tension with moral realism (H&T 2000: 139-142).

Notwithstanding H&T's contention, David Brink (2001) has recently advanced a novel account of content-fixing for moral terms that is promised to avoid the threat of an MTE counterexample. In the present chapter, I examine Brink's proposed moral semantics. I argue that his semantics fails to yield a solution to MTE that is compatible with naturalist moral realism. In particular, his semantics impales SEN on the horns of a dilemma. Understood one way, his semantics is incompatible with the stance-independence of putative moral facts; and thus, it is incompatible with moral realism. Understood another way, it requires the acceptance of non-natural facts, and so, is incompatible with ethical naturalism. Because SEN is a form of both moral realism and ethical naturalism, whatever solution Brink's moral semantics offers with respect to MTE is a solution that SEN cannot avail itself of.

4.2. Brink's Moral Semantics.

4.2.1. Brink's Moral Semantics: an initial formulation.

Brink's account of content-fixing for moral terms is presented in his "Realism, Naturalism, and Moral Semantics" (2001). One of the main attractions of this account is that it is supposed to provide the ethical naturalist with an answer to the argument from chauvinistic conceptual relativism. Although it is a form of semantic externalism, Brink's moral semantics re-emphasizes a role for speakers' intentions with respect to content-fixing. To illustrate the role of speakers' intentions with respect to content-fixing Brink begins with an example involving a non-moral, natural kind term. He maintains that, when speakers introduce such a term into their lexicon, their referential intentions determine which feature of our environment they are naming. For example,

> …those who introduced the term 'water' intended to refer to the structure, whatever it is, that explains the perceptible and functional features of the colorless, odorless stuff—found in lakes, rivers, etc.—that is suitable for drinking, bathing, and supporting life. It is this intention that fixes the reference of 'water'. As it turns out, it is the chemical microstructure H_2O that answers this explanatory description (Brink 2001: 172).

Thus, the referent of the kind name 'water' is the chemical kind H_2O, while, presumably, the content of the corresponding predicate 'water' is the property of *being H_2O*.

(It is important to note that, for Brink, the speakers' referential intention functions only to identify the content of 'water.' The content of the description embodied in their 'water'-related referential intention does not itself become the semantic content of 'water.' For if it did, then the property expressed by 'water' would be something like *being a colorless, odorless stuff found in lakes and rivers that is suitable for drinking, bathing and supporting life*. If this were the content of 'water,' then we would have to

say that Putnam's *XYZ* belongs to the extension of 'water'; but this is not what Brink wants to say.)

Turning his attention to moral discourse, Brink writes that

> …we need some parallel descriptive specification of the referential intentions of moral inquirers that would justify us interpreting a community of inquirers as engaged in moral inquiry…[B]ut it must be a description that is sufficiently abstract, so that a wide variety of views…might be thought to satisfy this description. Moreover, what best satisfies this description must be a matter of substantive moral theory (*ibid*.).

He ultimately recommends the following descriptive specification of the content-fixing intentions of moral inquirers:

> … we should understand perhaps all moral appraisers, and certainly those who introduced moral categories and terms, as using those categories and terms with the intention of picking out those properties of people, actions, and institutions—whatever those properties are—that play an important role in the interpersonal justification of people's characters, their actions, and their institutions (*ibid*. 174).

If Brink is right, then we should understand speakers on both Earth and Twin Earth as using 'right' (and t-'right) with the intention of expressing a property (or properties) that play "an important role in interpersonal justification." Only under this assumption are we permitted to view both groups as engaged in *moral* inquiry and as making *moral* judgments when they use their respective predicates. If it were to turn out that Twin Earthlings use t-'right' without the intention of picking out properties that play an important role in interpersonal justification, then we would have good reason to doubt after all that t-'right' is really translatable as our 'morally right.' Consequently, we would be within our rights to conclude that the resulting conceptual relativism is neither chauvinistic nor otherwise objectionable.

So far, so good. But if this suggestion is to succeed in steering ethical naturalism clear of chauvinistic conceptual relativism, then there needs to be some kind of guarantee

that all speakers who deploy seeming moral terms with this referential intention will succeed in fixing a common semantic content for those terms. How does Brink's semantics guarantee this? Here I quote Brink at length:

> Recall that relativism appeared to be a commitment of the theory of direct reference [i.e., semantic externalism] insofar as this theory was unable to identify a common meaning and reference about which appraisers from Earth and Moral Twin Earth held different beliefs. But our account of the shared referential intention to pick out people, actions, and institutions that are interpersonally justifiable, in virtue of which the judgments of Earthlings and Moral Twin Earthlings are both moral judgments, identifies just such a common meaning or reference about which the two communities have disagreement in belief. Their disagreement is one about which features of people, actions, and institutions make them interpersonally justifiable, with Earthlings holding consequentialist views and Moral Twin Earthlings holding deontological views. Moral realism and the theory of direct reference [i.e., semantic externalism] then, are compatible, and there is no reason to see a tension between ethical naturalism and moral realism (*ibid*. 174f).

Brink's explanation of how his semantics avoids chauvinistic relativism goes by too quickly. In order to see his solution more clearly, we will need to spell out his proposed semantics with more precision. To keep things simple, my formulation of his semantics will focus only on the deontic moral predicate 'right' as it applies to actions. For the same reason, I also omit mention of content-borrowing mechanisms that Brink discusses earlier in his paper.[1] Here, then, is one way to understand the externalist moral semantics that Brink proposes:

> BMS*: A predicate, 'F,' as used by the members of a linguistic community, C, is a deontic moral predicate translatable as the English 'morally right' and expresses the natural property N iff:
>
> 1. The members of C use 'F' with the intention of expressing a unique property of actions in virtue of which those actions are interpersonally justifiable for the members of C.

[1] Reference to such mechanisms may be unnecessary for our purposes since my formulation concerns itself with uses of a predicate by an entire linguistic community, as opposed uses by individual speakers.

2. An action, x, is (in fact) interpersonally justifiable for the members of C iff x instantiates N.

It is important to avoid confusion about what BMS* implies. The descriptive content of the referential intention to express a property that makes actions interpersonally justifiable serves only to fix a natural property as the semantic content of 'morally right.' The descriptive content of that intention does not itself serve as the content of 'morally right.' So, for example, if it should turn out that for the members of our community the property of *maximizing utility* makes acts interpersonally justifiable, then *maximizing utility* is the content and property expressed by our uses of 'morally right.' BMS* *should not* be understood as expressing the claim that the content of 'morally right' is the property *being interpersonally justifiable* or the property *being permitted by an interpersonally justifiable standard of action*.[2]

4.2.2. <u>A revised formulation of Brink's Moral Semantics.</u>

If BMS* is the correct interpretation of Brink's proposed moral semantics, then he has not supplied ethical naturalism with an acceptable answer to the problem of chauvinistic

[2] Two comments are worth mentioning here. First, this incorrect understanding of BMS* yields an account of the content of 'morally right' that is similar to the one defended by David Copp (1995). Very roughly, Copp's view is that 'morally right' expresses the property of *being permitted by a moral code that is justified for the action's circumstance* (*ibid*: 25f). Although Copp prefers to classify his view as a form of moral realism (*ibid*: 7, 223), in a recent paper, he acknowledges that, when his account is spelled out in detail, it does not satisfy the stance-independence requirement for the robust form of realism that we are interested in here (Copp 2005: 277).

A second point that merits comment is this: although I emphasize that BMS* implies that 'morally right' expresses (something like) the property of *maximizing utility* rather than (something like) the property of *being interpersonally justifiable*, this claim oversimplifies matters. If, as naturalists suggest (see §1.6.1), we are to view theories in first-order normative ethics as expressing property identities, then it is likely that we will need to assume that necessary co-extension is sufficient for property identity. However, given this assumption, it follows that if (e.g.) act-utilitarianism were true, we should probably conclude further that that *maximizing utility* is necessarily co-extensive with *being interpersonally justifiable*. And in that case, these properties are identical. From this, it follows that, given BMS* and the present assumptions, 'morally right' expresses *being interpersonally justifiable* after all. As best as I can tell, whether or not this really is an implication of BMS* (or an implication of the refined BMS that I present below) does not significantly impact the cogency of the arguments that I marshal against the view below. Thus, I will largely ignore this matter.

relativism. Let me explain. Given the description of the MTE scenario, it is natural to suppose that, for speakers on Earth, the property in virtue of which an act is interpersonally justifiable is the property of *maximizing utility*. It is also natural to suppose that, for speakers on Twin Earth, the property in virtue of which an act is interpersonally justifiable is the property of *treating no one as a mere means*. Unless these assumptions render the MTE story incoherent—and I see no reason to think that they do—it follows from BMS* that 'right' and t-'right' express different semantic content and cannot be used to engage in substantive moral disagreement.

It isn't difficult to see where BMS* goes wrong. BMS* attributes to moral speakers the intention to express the natural property that makes actions interpersonally justifiable *for their own community*. Moreover, the theory treats interpersonal justifiability a relation between actions and discrete linguistic communities. Because of these two features, it is possible that there be one natural property that makes actions interpersonally justifiable for Earthlings, and thus, serves as the content of 'right,' while another, different natural property makes actions interpersonally justifiable for Twin Earthlings, and thus, serves as the content of t-'right.' In other words, these two features render BMS* vulnerable to MTE counterexamples.

I do not believe that BMS* represents the moral semantics that Brink means to advance. However, this mistaken formulation of his view has revealed a crucial—though unspoken—assumption that underwrites his defense of naturalist moral realism: whatever natural property makes actions interpersonally justifiable, that property must make actions interpersonally justifiable for *all possible communities of moral agents*. A secondary assumption is that 'morally right' (and any other predicate translatable as

'morally right') is used by speakers with the intention of expressing whatever natural property makes an act interpersonally justifiable for *all possible* communities of moral agents.[3] It seems to me, then, that the moral semantics Brink means to propose is best captured by the following formulation:

> BMS: A predicate, 'F,' as used by the members of a linguistic community, C, is a deontic moral predicate translatable as the English 'morally right' and expresses the natural property N iff:
>
> 1. The members of C use 'F' with the intention of expressing a unique property of actions in virtue of which those actions are interpersonally justifiable for all possible moral agents.
>
> 2. An action, x, is (in fact) interpersonally justifiable for all possible moral agents iff x instantiates N.

This account of content-fixing for moral predicates appears to solve the problem of chauvinistic conceptual relativism. If BMS is true, then it is not possible for there to exist a community of speakers who use a predicate that is both translatable as 'right' and that expresses a natural property distinct from the property expressed by our own use of 'right.'[4]

It would seem, then, that Brink has supplied ethical naturalism with a moral semantics that eludes MTE-type counterexamples. In this respect, BMS represents a

---

[3] These two assumptions constitute very strong claims. I suspect something weaker might do the job for Brink's purposes. It might be argued on behalf of ethical naturalism that the more psychologically different from us that Twin Earthlings are, the less confident we are that t-'right' really does express the same content as 'right.' We are troubled only by MTE scenarios in which Twin Earth is populated with moral agents that fall within the "normal" range of human psychology. Fair enough. As long as it is granted that the Twin Earthlings described here fall within that range, I am happy to construe the locution 'all moral agents' as short for 'all moral agents *relevantly similar* to Earthlings.' I will, however, leave this qualification merely implicit in the reformulation of Brink's moral semantics that I am about to propose.

[4] In case this isn't clear: Suppose that, in fact, an action is interpersonally justifiable for all possible moral agents iff it instantiates *maximizing utility*. When BMS is conjoined with this assumption, it follows that any community of speakers who successfully use a predicate translatable as 'right' express *maximizing utility* by that predicate. If a community's use of 'right' failed to express *maximizing utility*, but expressed some other property instead, that would entail that they do not use 'right' with the intention of expressing a property that makes actions interpersonally justifiable for all possible moral agents. In that case, however, there is no pressure to view their predicate as translatable with our 'right.' Consequently, there is also no pressure to view their apparent disagreement with us as substantive.

significant improvement over Boyd's CSN. In the remainder of this chapter, I will argue that, whatever its merits as a semantics for 'morally right,' BMS is of no use to a sincere defender of realistic ethical naturalism. Upon closer inspection, it will be seen that BMS embodies commitments that are incompatible with a metaethics that is at once realist and naturalist.

4.3. Interpersonal Justification.

Brink makes no secret of the fact that the statement of his moral semantics makes essential use of a baldly normative term. He writes,

> [T]his account of moral semantics in terms of referential intentions to adopt the point of view of interpersonal justification is fiercely nonreductionist. To characterize the moral point of view in terms of interpersonal justification is to characterize it in ineliminably normative terms. This makes it a substantive question, which I have not addressed here, whether moral terms do refer and, if so, which properties they pick out (2001: 176).

As we see, Brink acknowledges that, by making the content of moral terms depend upon substantive normative facts about which properties make for interpersonal justification, BMS leaves us vulnerable to moral nihilism: If it were to turn out that there is no unique property that makes actions interpersonally justifiable for all moral agents, then 'morally right' would fail to express any property. In that case, all English sentences of the form 'ϕ is morally right' would be either false, or else lacking a truth-value.

The threat of nihilism from this direction is not trivial. But I want to set it aside for the moment (we will revisit it below). I am interested in a more subtle threat that arises for SEN as a result of the content-fixing role that BMS assigns to substantive normative facts about interpersonal justification. The threat is this: Judgments about interpersonal justification are themselves the subject of metaethical theorizing. If BMS is

to be utilized in a defense of ethical naturalism, then the metaethics that underwrites judgments about *interpersonal justification* had better be compatible with moral realism and ethical naturalism itself.[5] If it is not compatible, then BMS is in vain; its adoption would constitute an abandonment of naturalist moral realism, rather than a defense.

In order to evaluate the success of BMS, then, we need to examine what it is for an action (or for a standard of right-action) to be interpersonally justifiable. About this matter, Brink offers only a very broad sketch of an account. He writes that, according to the concept of morality that identifies moral standards with those that are interpersonally justifiable, "what is distinctive about the moral point of view is that we assess people, actions, and institutions according to standards that others can and should accept" (2001: 174). If by 'others' Brink means 'all moral agents,' his words suggest the following characterization of *interpersonal justification*:

> IJ: An action, x, is interpersonally justifiable iff x is permitted by a standard of right-action that every moral agent can and should accept.

With this conception of interpersonal justification in hand, I want to argue that ethical naturalism faces a dilemma when it is supplemented with BMS. To see the dilemma more clearly, I suggest one terminological adjustment to IJ. Instead of speaking of standards that agents *should* accept, I propose that we speak of the standards that agents have *normative reason* to accept.[6] By making this terminological swap, we can now avail ourselves of the distinction between internal and external normative reasons.

---

[5] Or, to be more precise: the metaethics that correctly accounts for judgments and facts about *interpersonal justification* had better not combine with BMS in a way that undermines either moral realism or ethical naturalism.

[6] I follow John Broome first in treating 'should' as roughly synonymous with 'ought,' and second in taking claims about what an agent ought to do as being equivalent—even if not synonymous with—claims about what an agent has "perfect" reason to do. Perfect (or all-things-considered) reasons differ from "*pro-tanto*" (or *prima facie*) reasons in that it is possible for an agent to have *pro-tanto* reason to ϕ even if it is false that he ought to ϕ. Such a case can arise when there is a more weighty *pro-tanto* reason not to ϕ. See Broome (2004: 34-42).

Roughly, an agent has internal reason to perform an action ϕ just in case ϕ-ing would contribute to the satisfaction of her (informed) desires (or other elements in her "subjective motivational set").[7] By contrast, external reasons obtain independently of agents' desires so that an agent can have external reason to ϕ, even if ϕ-ing would serve none of her informed desires. (I understand the distinction between internal and external reasons to be exhaustive).

In outline, the dilemma facing BMS is this: Suppose that IJ is read in such a way that judgments about which standard an agent should accept are understood as judgments about what an agent has *internal* reason to accept. Two problems ensue. The first is that it is doubtful that there is a single standard of action that all moral agents have internal reason to accept. If there is no such standard, then BMS entails moral nihilism. The second problem is that facts about what internal reasons an agent has are stance-dependent. When the internal reasons reading of IJ is combined with BMS, the stance-dependence of normative (internal) reasons is transferred to the moral facts themselves. Thus, the resulting metaethic is incompatible with moral realism.

For the second horn of the dilemma, we suppose that IJ is read in such a way that judgments about which standard an agent should accept are understood as judgments about what an agent has *external* reason to accept. Now, facts about what there is external reason to do are typically held to be irreducible to facts picked out by a purely non-moral vocabulary. Thus, the acceptance of such reasons appears to require that we

[7] The canonical statement of the distinction between internal and external reasons is Williams (1980). I will understand 'desire' in a broad way so as to include all the elements Williams includes in an agent's subjective motivational set. These include "dispositions of evaluation, patters of emotional reaction, personal loyalties, and various projects…embodying commitments of the agent" (*ibid*.: 105). It may be possible to offer a unifying account various phenomena included in a subjective motivational set by appeal to the notion of "direction of fit." Whereas belief is a state of mind that aims for its content to fit the world, elements of an agent's subjective motivational set (i.e., her desires) are states of mind that aim to get the world to fit their content.

expand our ontology. The trouble is, facts about what there is external reason to do fail to satisfy EC; they play no ineliminable role in our best available *a posteriori* explanations. This might lead some to want to opt for a reductive account of external reasons. But even if a reductive account could be made plausible, it appears that the resulting version of BMS is once again vulnerable to MTE and the problem of chauvinistic conceptual relativism.

Since the internal reasons interpretation and external reasons interpretation of IJ are exhaustive, it turns out that there is no conception of IJ that combines with BMS so as to solve the problem of chauvinistic conceptual relativism in a way that is consistent with both moral realism and ethical naturalism. Thus, BMS fails in its task: it is of no use in defending naturalist moral realism. In the remaining sections, I present the details of this dilemma, taking each horn in turn.

4.4. First Horn: Internal Reasons.

4.4.1. Internal reasons and the failure to converge.

On standard versions of reasons internalism (the view that all genuine normative reasons are internal reasons), the desires relevant to an agent's having a reason to ϕ are taken to be those desires that she would have, were she in ideal epistemic circumstances – e.g., in a condition of full information and correct deliberation (cf. Williams 1980; Smith 1995). Call the desires that an agent has under such idealized circumstances her *ideal desires*. For a reasons internalist, whether an agent has reason to accept a given standard of right-action depends upon whether her accepting that standard will satisfy her ideal desires. The trouble for BMS on the internal reasons-reading of IJ, is that it is very unlikely that a

single standard of right-action serves the ideal desires of every individual moral agent. For example, a very strong, clever, and unsympathetic agent with Nietzschean tastes would probably be better served by accepting a standard different from the standard appropriate for someone with a slower wit and a more delicate physique and sensibility. In that case, however, there is no standard that all agents have reason to accept. Given IJ, it follows that no actions are interpersonally justifiable. It follows further that BMS implies our use of 'morally right' expresses no property. We are left with moral nihilism.

4.4.2. Smith's absolutist conception of internal reasons.

This is too quick, however. Some philosophers reject this relativistic conception of internal reasons in favor of an absolutist conception. Perhaps the ethical naturalist can escape the present difficulty by drawing on their work. In this section, I will consider what the ethical naturalist might gain by adopting the sort of absolutist conception of internal reasons advanced by Michael Smith.

According to Smith's conception of an internal reason, an agent counts as being in ideal epistemic circumstances only when her entire set of desires is "systematically justified" in the sense that her desire-set is brought to exhibit maximal coherence and unity. The process of justification among desires that Smith envisions is akin to Rawls's description of the method for achieving a reflective equilibrium among one's beliefs (1971/1999: 40-46): we begin with our actual stock of desires and then add and subtract desires until we reach a set of desires that is maximally coherent and unified. Call an agent *fully rational* iff her desires are systematically justified, she has no false beliefs, and she has all relevant true beliefs. Smith's view is that an agent has an internal reason

to perform ϕ in circumstance C "if and only if, if she were fully rational, she would desire that she ϕs in C" (1995: 112). Importantly, Smith believes that, under conditions of full rationality, "all possible rational creatures would desire alike as regards what is to be done in the various circumstances they might face…" (*ibid*. 118). If he is right, then his account allows for a form of absolutism about internal reasons. An implication of this account is that there will be no variation among possible agents with respect to the matter of which standard of right-action they have internal reason to accept.[8] It would seem, then, that BMS can evade the threat of moral nihilism if we incorporate an absolutist version of internal reasons into our reading of IJ.

The trouble with this strategy is that it can be effective only if, in fact, there really would be a convergence in the desires of *all possible* fully rational creatures. I do not share Smith's optimism that such a convergence is forthcoming. It strikes me as a genuine possibility that there could be a moral agent whose informed desires achieve maximal coherence and yet they are different from the desires of other fully rational moral agents with respect to matters of moral importance. If it is a genuine possibility that there be one such agent, then I see little reason to deny that it is a genuine possibility that there could be an entire planet populated with billions of similar agents.[9] If this is possible, then the internalist reading of IJ entails that there is no standard of right action

[8] Perhaps I am being too generous towards BMS. Even granting Smith's absolutist internalist account of normative reasons, it does not follow that all agents will have normative reason to accept the same moral standard. This is because Smith's account still relativizes the reasons an agent has to her circumstances. (What make it absolutist, nevertheless, is that she has reason to ϕ only if all ideal agents would agree in their desire that, given her circumstances, she ϕs). It is compatible with Smith's account that, given differences in the circumstances faced by Earthlings and Twin Earthlings, members of each group respectively have reason to accept a different standard of right-action.

[9] Horgan and Timmons press this point themselves in their own critique of Smith's theory of reasons (1996b: 210-211). They note that a defender of Smith's account must reject such a possibility as "misdescribed" or, in any case, illusory. For what it's worth, in a review of Thomas Carson's *The Status of Morality*, Brink raises what is essentially the same objection against Carson's ideal observer theory of moral facts that I am raising against Smith here (1986: 146).

that is interpersonally justifiable for all moral agents. Once again, the internalist theory of normative reasons conjoins with BMS to entail moral nihilism.

4.4.3. Internal reasons and the stance-dependence of normative facts.

To avoid the present difficulty, the naturalist moral realist must insist that the agents I have just described are not genuinely possible. A reply along these lines might have some plausibility if there were a "normative reality" existing independently of the desires we actually have or would come to have via the method of reflective equilibrium. If there were a single, independent, normative reality, then the naturalist could argued that, for every possible agent, contact with this reality constrains and guides the process of systematically justifying her desires. Because each agent's process of desire-justification is guided by a single normative reality, it may be argued further that, insofar as the process of justification is properly implemented by each agent, it is reasonable to believe that all agents would converge in their ideal desires. If all of this could be maintained, then I am ready to grant that there would at least be *some* grounds for dismissing as mistaken the intuition that it is possible that there be a fully rational agent whose desires are importantly different from those of other fully rational agents.

Whatever the merits of the defense just sketched, it is not available to a proponent of reasons internalism. For Smith, as for other defenders of reasons internalism, there simply is no antecedently existing normative reality to guide the process of desire-justification undertaken by any given agent. Instead, the facts about what there is normative reason for an agent to do are *constituted* by facts about which set of desires she would have, were she successfully to complete the process of systematic justification

among her fully informed desires (along with the non-normative facts about which actions would satisfy those desires). The trouble for the absolutist version of reasons internalism, then, is that in the absence of an independently existing normative reality, it is hard to see a reason to accept that it is a necessary truth that all moral agents who subject their desires to the process of justification that Smith describes end up desiring exactly alike.

The preceding observations prepare us for the second major problem that arises for BMS when IJ is given an internalist reading: If, as reasons internalists claim, facts about what an agent has normative reason to do are themselves partly constituted by facts about what that agent would desire, were she fully rational, then having a normative reason must be regarded as a stance-dependent kind of fact. Because of this, reasons internalism must be classified as a form of constructivist anti-realism about normative facts. In and of itself, this need not entail constructivism about *moral* facts, since one might deny that moral facts are a species of normative facts. However, because, according to BMS, facts about which natural properties are morally right-making are determined by facts about which moral standards agents have normative reason to accept, the stance-dependence of the normative facts spreads to the moral facts themselves. Let me try to spell out more clearly why this is.

When the internalist reading of IJ is conjoined with BMS, the matter of which property 'morally right' expresses is made dependent upon facts about which standard of right action all moral agents would desire that they accept, were they fully rational. Because of this, the standard that rational agents would accept is *ipso facto* the true first-order theory of *moral rightness*. This is because this very standard determines which

natural property is expressed by 'morally right' and thus, which natural property makes right acts right. For example, given this reading of BMS, the property of *maximizing utility* will make right acts right if (and only if) all agents would desire that they accept act-utilitarianism as their moral standard, were they fully rational; and, as a result, AUh would be the true theory in the normative ethics of behavior. On the other hand, if all agents would desire that they accept a form of Kantianism, were they fully rational, then (something like) the property of *treating no one as a mere means* would make right acts right instead; and in that case, CI-2 would be the true theory in the normative ethics of behavior.

It should be obvious that the internal reasons way of understanding BMS has put us in a situation in which the moral standard that fixes the facts about which acts are morally right is made true in virtue of its being ratified from within the perspective of hypothetical, fully rational agents. In other words, moral facts fail to be stance-independent on this account.[10] Consequently, BMS, when combined with an internal reasons reading of IJ, cannot be deployed in defense of naturalistic moral realism. Under the most optimistic assumptions, the most robust metaethic that it can support is an absolutist version of moral constructivism.

4.5. Second Horn: External Reasons.

4.5.1. BMS, IJ, and external reasons.

In light of the troubles raised in the previous section, it looks like a successful defense of ethical naturalism using BMS will have to adopt an external reasons interpretation of IJ. On this interpretation, an act is interpersonally justifiable iff it is permitted by a standard

[10] The stance-independence requirement was discussed above, in §1.2.1.

of right-action that every moral agent has external reason to accept. An agent's external reasons, recall, are roughly those normative reasons that an agent has independently of her desires (be they actual desires or idealized desires).[11] Because of this independence, the version of BMS that incorporates an external reasons interpretation of IJ is not obviously vulnerable to the charge that it renders moral facts stance-dependent. In this respect, the reasons externalist version of BMS looks to be compatible with moral realism.

Unfortunately, by committing BMS to the existence of external reasons, this interpretation of IJ raises troubles of its own. Note first that the internalist account of reasons considered in §4.4 is a reductive view: facts about what there is internal normative reason to do reduce to facts about what agents would desire under actual or counterfactual circumstances. Provided that desires and modal facts are open to empirical investigation, facts about what there is internal reason to do are natural facts.[12] By contrast, few have claimed that facts about what there is external reason to do are similarly reducible.[13] It seems, then, that a commitment to external reasons requires us to expand our ontology in a way that a commitment to internal reasons does not. The worry, however, is that a commitment to irreducible external reasons expands our

[11] More precisely, an agent's external reasons are reasons she has that are not *constituted* by anyone's actual or ideal desires. Strictly speaking, a reasons externalist can allow that some external reasons have a certain kind of dependence on desires: we may have external reasons to perform actions that would disappear if we lost certain desires. For instance, it may be that Beth has a reason to see a horror film, but only if she desires to see a horror film. If she loses her desire to view the film, she loses her reason to view it as well. Even so, this reason can still be external insofar as it is grounded in an even more fundamental external reason. For example, it may be that every agent has external reason to fulfill her own desires. Notice that this latter reason need not be thought of as dependent upon the agent's desires: we might have reason to fulfill our desires even if we would not (ideally) desire that we fulfill our desires.

[12] I expect that some readers will be unsympathetic to the idea that modal facts are open to empirical investigation. I cannot address this concern here; nor is there need to address it here. If the modal facts that underwrite internal reasons cannot be made to fit within an empirical epistemology—either by being shown to be analytic or else knowable via inference to the best *a posteriori* explanations—then so much the worse for ethical naturalists.

[13] Nevertheless, in §4.5.3 I consider the prospects for BMS given a reductive view of external reasons.

ontology beyond what can be discovered by solely empirical means. If so, then external reasons cannot be accommodated within a naturalistic ontology. In that case, BMS will have purchased the stance-independence of moral facts at the price of abandoning ethical naturalism.

4.5.2. Naturalism and external reasons.

Is there anything to the worry that irreducible external reasons cannot be naturalized? I think there is. Recall the account of *natural propertyhood* from §1.3.1:

> NP: a property, P, is natural just in case a synthetic proposition to the effect that an individual instantiates P can be known only by way of empirical investigation, if it can be known at all.

In light of NP, the property of *being an external reason* is a natural property only if knowledge that some consideration is an external reason can be acquired and justified using solely empirical methods (assuming, of course, that any such knowledge can be attained at all).

How plausible is it that knowledge of external reasons is empirical? I think it is not very plausible. In the first place, it is obvious that we do not detect external reasons through direct sensory perception. (To put the same point more carefully: we do not detect that some considerations [i.e. states of affairs] are external normative reasons through the use of our five faculties of sense-perception—sight, hearing, touch, taste, and smell). It may be worth noting a feature of reasons-talk that is apt to cause confusion on this point. The thing that we often identify as the normative reason to ϕ often is something (in particular, a state of affairs) that is empirically knowable. For instance, we might identify the fact that Carl is drowning as the reason I ought to throw him a life

preserver. Here, the fact that Carl is drowning may plausibly be thought of as something we can observe with our senses.[14] But the fact that Carl is drowning is not the sort of fact that we are concerned with presently. What is at issue, rather, is how we come to recognize the observed fact that Carl is drowning *as a reason* to throw the life preserver. To put it another way: we are interested to discover how we come to know that Carl's drowning makes it the case that we (*pro-tanto*) ought to throw him the life preserver. I contend that we do not discover this sort of fact by way of sensory perception.

Moreover, because we do not perceive instances of something's being an external reason, it should be obvious that we cannot arrive at the judgment that we have reason to ϕ by way of enumerative induction from our sensory observations. The only remaining empirical way of discovering which states of affairs are external reasons, then, is by way of an inference to the best explanation of our empirical observations (i.e., abduction).[15] But what sorts of observable phenomena might external reasons help to explain? Certainly, such reasons are of no use in explaining phenomena such as planetary motion, geological processes, the life-cycles of plants and bacteria, the behavior of subatomic particles, etc.; or, more precisely, external reasons are no part of any plausible explanation of such phenomena unless those phenomena are themselves caused by the behavior of intentional agents. (Note that to deny this would be effectively to accept a teleological view of the natural world.) In fact, I believe that the only sort of phenomena

[14] Of course, as with any case of perception, the ability to observe that someone is drowning is theory laden. In this case the perception presupposes background knowledge of mammals, respiration, mortality, etc.

[15] I have not said anything about *a priori* conceptual analysis. Some might contend that conceptual analysis is an empirical way of knowing; and so, it might be thought that I ought to address whether we can come to know what we have external reason to do by this method. This sort of complaint misses the mark. Our present question is how we come to know the existence of *irreducible* external normative reasons. Thus, the possibility that we could naturalize a certain kind of external reason by conceptually reducing it to some kind of less problematic phenomena is not relevant here. (I consider the prospects for BMS given a reductive account of external reasons in §4.5.3 below).

for which it is even remotely plausible that external reasons play an explanatory role are the beliefs and intentional actions of agents.

Before considering the prospects of explaining beliefs and actions by appeal to external reasons, I want to take a moment to point out that this task is of double importance for the project of naturalizing external reasons. Presently, we are pursuing the question of whether knowledge of the (putative) external reasons that we have is empirical knowledge. An affirmative answer is needed if we are to construe external reasons as natural properties, in accordance with NP. But finding ineliminable work for external reasons to perform in our best *a posteriori* explanations is also important for the naturalist's project because of his commitment to the ontological criterion EC, presented in §1.3.2:

> EC: posit the existence of an entity (or a kind of entity) if and only if reference to that (kind of) entity is needed in our best available *a posteriori* explanations of observable phenomena.

Unless it can be successfully argued that irreducible external reasons play an ineliminable role in the best explanation of the beliefs and intentional actions of agents, then, reasons externalism will fail ethical naturalism twice over.

But are external reasons needed in the best explanations of our beliefs and intentional actions? To simplify, we can set aside discussion of intentional actions and focus on whether external reasons are needed to explain certain beliefs we hold—in particular, whether they are needed to explain those beliefs whose contents are propositions to the effect that there are external reasons to perform certain actions (call beliefs of this sort, *normative beliefs*). The reason I say this is because paradigmatic cases in which an external reason plausibly forms part of the explanation of an agent's

intentional act of ϕ-ing will be cases in which an agent ϕed because *she believed* that she had external reason to ϕ (or, what amounts to roughly the same thing, because she believed that she ought to ϕ). This is not to insist that there couldn't be cases in which an external reason bypasses an agent's beliefs and by some subconscious process causes her to intend to ϕ. But it seems to me that if external reasons are never part of a conscious process that results in an agent forming an intention to act, it is doubtful that we will find compelling examples of actions where no better explanation can be found than ones that posit external reasons operating subconsciously. In addition, we should expect that false external reasons beliefs can bring about intentional action just as effectively as true external reasons beliefs. If so, then all we can learn from an agent's action is that she believed herself to have a reason; but it does not follow from this that she really did have a reason.

The present challenge facing the naturalist is to show that external reasons are part of the best *a posteriori* explanations of our (or anyone's) normative beliefs. Above, I argued that knowledge of external reasons—if any such thing is to be had—does not arise from direct perception; nor does it arise from inductive inference; nor does it arise by way of abductive inference from observations of inanimate objects. As we saw, the only remaining phenomena that might be explained by external reasons are agent's normative beliefs (and the actions they give rise to). So perhaps it can be argued that we come to know the existence of external reasons because we must posit them in order to explain the normative beliefs of agents. But with this, we have returned to our original question; we have made no progress.

I think the conclusion to draw here is that our normative beliefs do not have their source in any empirical knowledge-gathering processes. On the face of it, this would seem to decertify external reasons as natural properties or facts in light of NP. But perhaps this is too quick. Even if our normative beliefs do not have their *source* in empirical processes, it may still be the case that they can be epistemically *justified* using empirical methods (cf. Devitt 1999: 46). If so, then a non-empirical etiology for normative beliefs need not be incompatible with the commitments of ethical naturalism.

If our normative beliefs do not have their source in empirical processes, then where do they come from? Two answers suggest themselves. First, it may be that we arrive at our normative beliefs by some kind of synthetic, *a priori* intuition through which we "grasp" that some considerations are external reasons for acting. Such a process, so described, is incompatible with naturalism, as it renders external reasons non-natural according to NP. A more promising answer for the naturalist is that normative beliefs are innate. As far as I am aware, the only plausible account of innate beliefs that is compatible with naturalism is an evolutionary account.[16] According to an account of this sort, a disposition to make normative (external reason) judgments arose in our species by way of natural selection. Presumably, those who press this account must argue that having a normative sensibility (i.e., a disposition to make normative judgments) enhanced the reproductive fitness of our ancestors.[17]

---

[16] Examples of competing hypotheses that are not so compatible with naturalism include the hypothesis that God implants these beliefs in us and the Platonic hypothesis that we recollect these truths from past lives.
[17] Another possibility is that the disposition for normative judgment is a "spandrel," a mere accidental by-product of some different trait that was selected for. But notice that if this disposition is a spandrel, then there isn't any reason to suppose that real external reasons played a role in shaping its development; and if that is so, then external reasons are not needed in order to best explain the development of our innate disposition to make normative judgments and have normative beliefs. Since external reasons aren't needed to explain anything else, this fact along with EC implies that naturalists ought to deny their existence.

I contend that even if normative beliefs are innate, this hypothesis does not bode well for the claim that external reasons are natural facts. In Chapter 6, I discuss an evolutionary account of moral judgment in some depth. According to that account, putative moral facts play no ineliminable role in the best explanation of our making the sorts of moral judgments that we make. The very same account can also be extended to offer an explanation of our normative, external reason judgments in general. I refer those looking for a more detailed discussion of the evolution of normative judgment to that chapter (see especially, §§6.3.2 - 6.3.4). Here, I will briefly indicate why the evolutionary account of external reason judgments will not be of any help for naturalists looking to justify their acceptance of external reasons.

The trouble is that external reasons themselves (if there are any such things) play no ineliminable role in the most plausible evolutionary account of normative judgment. Normative beliefs need not be accurate—that is, they need not accurately represent some putative normative reality populated with external reasons—in order to enhance the reproductive fitness of creatures that make such judgments. Let me explain. To begin with, any plausible evolutionary account of normative judgment will have to suppose that there is a very strong (even if contingent) connection between judging that one has a reason to ϕ and being motivated to ϕ. Without an assumption of this sort, the making of normative judgments (be they accurate or inaccurate judgments) may not be able to influence behavior to a great enough extent that it enhances reproductive fitness. Now consider a human (or at least, a hominid) that is disposed to judge that the fact that something is her offspring is a strong reason to feed and care for it. It is more plausible than not that, in the sort of environment in which our ancestors lived, the disposition to

make this sort of judgment (and to have one's actions be guided by it) enhanced the reproductive fitness of the creatures that had it. By contrast, we would expect to find a lower degree of reproductive fitness for a similar human who instead judges the fact that something is her offspring to be a reason to ignore it. Notice, however, that we would expect the first creature to enjoy greater reproductive fitness than the second *even if it were not true that something's being one's offspring is a strong reason to feed and care for it*.

The upshot, then, is that a creature's normative beliefs can influence her to behave in fitness-enhancing ways even if those beliefs are inaccurate. Because of this, it is hard to see what ineliminable role external reasons can play in the story of how our ancestors evolved to have innate normative beliefs.

To conclude, irreducible external reasons cannot be accommodated by ethical naturalists. In the first place, there is no credible account of our knowledge of such reasons that is compatible with a commitment to empiricism. Consequently, external reasons fail to be natural properties given NP. In addition, external reasons explain no observable phenomena; nor do they even explain the fact that we have beliefs with external reasons as part of their content. Thus, given the naturalist's ontological criterion, EC, naturalists cannot countenance irreducible external reasons. Because of this, the irreducible external reasons reading of IJ cannot be combined with BMS to yield a semantics that honors the naturalistic metaphysical commitments of SEN.

4.5.3. Reductive accounts of external reasons.

What are the prospects for BMS if we incorporate a *reductive* account of external reasons? To examine the prospects of a reductivist maneuver, we will need some sample reductions of normative reasons. Those reductivists who favor a hedonistic version of rational egoism might be inclined to accept something like the following reductive account:

> (REh) the property of *an agent's having external reason to* ϕ = the property of ϕ *maximizing its agent's hedonic utility*.

I grant that REh renders external reasons acceptable from the point of view of metaphysical naturalism. I am also ready to grant that REh satisfies (or could be made to satisfy) the stance-independence requirement. Consequently, I accept (or at any rate, I do not deny) that REh could serve as a robustly realist, naturalistic account of normative reasons.

Although I have several misgivings about reductive accounts along the lines of REh, I will mention only one. The naturalist who advances REh owes us a semantics of the predicate 'has a normative reason.' If that semantics is a version of descriptivism, then it will be easy to devise a "Normative Twin Earth" thought experiment that will reveal that naturalism is now committed to a chauvinistic conceptual relativism about 'has a normative reason.' We need to imagine only that Earthlings subscribe to a hedonistic form of rational egoism while Twin Earthlings subscribe to a perfectionist form of rational egoism. A similar result arises if it is suggested that we adopt a causal semantics for normative terms instead. In that case, we could imagine a scenario in which Earthling uses of 'has a normative reason' are causally regulated by the property *maximizing the agent's hedonic utility*, while Twin Earthling uses of 'has a normative

reason' are causally regulated by the property *maximizes the degree of perfection of the agent's essence*. Once again, Boyd's causal semantics entails chauvinistic conceptual relativism.

To deal with this difficulty, we might try to mimic Brink's moral semantics here and say that the content of 'has a normative reason' is fixed in virtue of speakers' using this predicate with the intention of picking out a natural property that plays an important role in *personal* justification for all possible agents. But here we are again faced with the worry that there may not be any such property, in which case, normative nihilism results. Moreover, even if there is such a natural property, we now need a metaethical account of judgments about personal justification. But this puts us back where we started. Even if 'personal justification' successfully denotes a real property, that property could turn out to be incompatible with metaphysical naturalism, or it may turn out to combine with the new normative semantics in such a way as to yield constructivism about normative reasons. In fact, I believe that this semantics for 'having a normative reason' faces exactly the same dilemma faced by BMS. In short, a reductive account of external reasons does not answer the worries facing BMS; at best, it merely pushes them back another step.

4.6. Conclusion.

Brink has proposed a moral semantics whereby the content of a predicate such as 'morally right' is fixed according to whichever standard of conduct moral agents can and should accept. This semantics promises to avoid the problem of chauvinistic conceptual relativism while preserving an account of moral facts and properties that is both realist

and naturalist. I have argued that Brink's proposal cannot satisfy all three desiderata at once. If the reasons we have for accepting a standard of conduct are internal reasons, then the moral metaphysics that results from Brink's semantics is at best a form of moral constructivism and at worst a form of moral nihilism. On the other hand, if the reasons we have for accepting a standard of conduct are external reasons, ethical naturalism cannot be sustained. I conclude that Brink's externalist moral semantics fails to supply naturalistic moral realism with a satisfactory answer to the problem of chauvinistic conceptual relativism.

CHAPTER 5

IS GOODNESS A HOMEOSTATIC PROPERTY CLUSTER?

5.1. Introduction.

In Chapter 1, we saw that synthetic ethical naturalists adopt an externalist semantics for moral predicates in order to respond to a battery of objections to traditional versions of naturalistic moral realism. The primary benefit of this kind of moral semantics is that it makes it possible to view naturalistic definitions of moral predicates (and statements of identity between moral and natural properties) as expressing putative synthetic, *a posteriori* necessities. In Chapter 2, I presented Horgan and Timmons' Moral Twin Earth argument against moral semantic externalism. There, and in Chapters 3 and 4, I defended their argument from a number of important replies. If my defense has been successful, then we appear to be justified in concluding that the MTE argument refutes moral semantic externalism and SEN along with it.

Suppose, however, that the MTE argument falls short of a total refutation of SEN. Even in that case, I would insist at the very least that the MTE argument changes the dialectical situation with respect to SEN. In the past, when confronted with misgivings about the supposed analyticity of moral identity claims, ethical naturalists have thought it sufficient simply to note the existence of non-analytic, *a posteriori* identities and definitions that are related to natural kinds and natural kind terms that fall within the purview of the natural sciences. It was presumed that, if it is an *a posteriori* matter that *being water* is identical with *being H2O*, there is no reason to deny that it is an *a posteriori* matter that *moral rightness* is identical with (e.g.) *maximizing hedonic utility*. I

believe that the Moral Twin Earth argument shows that this presumption is mistaken: the proponents of ethical naturalism need to offer compelling, independent reasons for thinking that the semantics and epistemology appropriate for natural kinds and natural kind terms ought to be extended to moral properties and moral predicates.[1]

In "How to Be a Moral Realist," Boyd offers a novel account of moral properties that has the potential to provide some of the independent justification that is needed for a defense of externalist moral semantics and SEN. He suggests that some moral properties have the same metaphysical structure as properties that define natural kinds. In particular, he proposes that *moral goodness* is constituted by a "homeostatic property cluster" (HPC). As we will see below, the properties in an HPC are unified by nomological necessity, not by conceptual necessity. Because of this, if an HPC serves as a kind's essence, then the question of which properties belong to that kind's essence can be answered only by discovering which properties exhibit the right nomological connection to the rest of the clustered properties. This, however, is an *a posteriori* question, not to be answered by way of *a priori* conceptual analysis. Thus, Boyd writes: "If the good is defined by a homeostatic phenomenon the details of which we still do not entirely know, then it is a paradigm case of a property whose 'essence' is given by a natural [i.e., *a posteriori* real] rather than a stipulative [i.e., analytic or nominal] definition" (Boyd 1988: 210). In this way, Boyd's HPC conception of *moral goodness*, if viable, promises an important independent justification for the central semantic and epistemological claims of SEN.

---

[1] At least one ethical naturalist seems to agree that independent justification is needed. Boyd writes that, if the naturalistic moral realist is to legitimately make use of the epistemological and semantic claims characteristic of synthetic ethical naturalism, then there needs to be "good reasons to think that moral terms must possess natural [i.e. non-analytic, *a posteriori*] rather than stipulative [i.e., analytic or nominal] definitions" (1988: 210, cf. 201).

In this chapter, I argue that Boyd's hypothesis is false: *moral goodness* is not an HPC. In §5.2, I present Boyd's account of HPC kinds. In §5.3, I present his proposal that *moral goodness* is constituted by an HPC (and thus demarcates an HPC kind). In §5.4, I advance two arguments against this proposal. The first is a moral argument. The second points to suspicious structural features of THE MORAL GOOD[2] that are not shared by paradigmatic HPC kinds. In §5.5, I offer two further arguments to the effect that reference to THE MORAL GOOD does not support reliable inductive inference in the way that it should were it an HPC kind. In §5.6, I anticipate a reply that might be made on behalf of the HPC conception of *moral goodness*.

5.2. Boyd's Homeostatic Property Cluster Kinds

5.2.1. Homeostatic Property Clusters.

Since Mill's *A System of Logic*, philosophers have witnessed an intimate connection between natural kinds, on the one hand, and induction and explanation on the other hand (Mill 1867: 434; cf. Boyd 1999b: 81; Dupré 1981: 68; Kitcher 1984: 315n; LaPorte 2004: 19; Quine 1969: 126; Russell 1948: 318). Boyd writes: "One of the defining features of natural kinds generally…is that reference to natural kinds facilitates induction and explanation with respect to a wide variety of issues" (1999b: 81). Hilary Kornblith adds: "It is precisely because the world has the causal structure required for the existence of natural kinds that inductive knowledge is even possible" (1993: 35; cf. Millikan 2000: 15-32).

[2] In this chapter, I employ a convention of using small caps for terms referring to kinds and continue to use italics for terms referring to properties. I have found that treating properties as distinct from the kinds whose membership they define helps in the exposition of Boyd's view. Nothing I say in this chapter depends upon kinds and properties actually being distinct sorts of entities, however.

Boyd's HPC account of natural kinds is intended to explain, among other things, how it is that certain natural kinds ground induction and explanation. The key idea is that, for many natural kinds, the essence of the kind is constituted by a group of properties that, although logically independent of one another, are "clustered in nature in the sense that they co-occur in an important number of cases" (Boyd 1988: 197). The clustering is the result of a certain nomological relationship among the properties. Boyd describes this relationship as a "sort of homeostasis." He offers a disjunctive account of what is involved in this kind of homeostasis for a family of properties, F:

> Either the presence of some of the properties in F tends (under appropriate conditions) to favor the presence of the others, or there are underlying mechanisms or processes that tend to maintain the presence of the properties in F, or both (*ibid*.).

Two things deserve mention here. First, it is plausible to read the phrase 'tends to favor' as meaning "makes more likely." Making this substitution naturally raises the question of just how much more likely need the presence of some of the properties make the presence of the others in order for there to be a homeostatic clustering of properties. Would any increase in likelihood, however slight, suffice for a group of properties to count as homeostatically clustered? Although Boyd does not explicitly set a lower bound on how much of an increase in likelihood is required for a group of properties to count as homeostatically clustered, he does make it clear that, where HPC kinds are concerned, the homeostatic clustering of properties should be "causally important." This suggests that, at least with respect to HPC kinds, we should expect that the presence of some properties in the relevant group raises to a considerable degree the likelihood of the others being present. (Indeed, if this were not so, then it would be hard to see how HPC kinds could fulfill their role in facilitating reliable inductive inference.) Of course, even if this is

accepted, we are still left with a good deal of imprecision in the account of property homeostasis. Although I cannot attempt to sharpen the account any further, I should note that this imprecision is relevant to the discussion in §5.6 below.

The second item worthy of mention is this: The passage quoted above suggests that the mechanisms responsible for the homeostasis among a family of properties should be thought of as "underlying" in some sense. However, in other writings, Boyd makes it clear that, in some cases, (some of) the relevant mechanisms responsible for property homeostasis are external to the individual members of the HPC kind (Boyd 1999b: 79). With this in mind, I propose that we drop the word 'underlying' from Boyd's account of property homeostasis.

In light of these considerations, I take the following as my official statement of Boyd's account of property homeostasis:

PH: A family of properties, F, is homeostatically clustered if and only if either:
- (i) (under appropriate conditions) the presence of some of the properties in F makes more likely the presence of the other properties in F, or
- (ii) there exist mechanisms or processes that make more likely the continued presence of the properties in F, or
- (iii) both i. and ii.

5.2.2. Homeostatic Property Cluster Kinds.

For Boyd, some HPCs constitute the real essences[3] of certain natural kinds.[4] Call such kinds 'HPC kinds.' In their capacity as essences, HPCs (along with the mechanisms that

[3] The real essences of kinds are contrasted with "nominal" essences. Roughly, the nominal essence of a kind, K, is something like an analytic definition that speakers conventionally associate with the predicate or kind term that corresponds to K. If a kind has only a nominal definition, then its membership conditions depend solely upon linguistic conventions and are discoverable by *a priori* conceptual analysis. By contrast, the real essence of a kind determines that kind's membership conditions independently of linguistic conventions and thus cannot be discovered by mere conceptual analysis (cf. Boyd 1988: 194f; 1999a: 142, 146; Ellis 2001: 32). (Note that Boyd uses 'real essence' interchangeably with 'natural definition' and '*a posteriori* definition'.).

bind them together) supply the membership conditions of HPC kinds. As I understand Boyd, an individual, x, is a member of an HPC kind K just in case (a) x instantiates the properties in the HPC that defines K and (b) the co-instantiation of these properties in x is brought about or maintained (at least in part) by the homeostatic mechanisms definitive of K (if there are any such mechanisms).[5]

Homeostatic property cluster kinds are well suited to satisfy the inductive and explanatory role that is associated with natural kinds. Because the defining properties of an HPC kind are homeostatically unified, the members of a given HPC kind will exhibit a significant degree of uniformity. In turn, this uniformity facilitates reliable inductive inferences and explanations. This sort of uniformity is readily seen, for example, in biological species. For an individual member of a given species, its manifest properties flow from underlying mechanisms. Because these mechanisms are shared by every (or

---

[4] In comments on an earlier version of this chapter, an anonymous editor for *Ethics* recommended a different reading of Boyd with respect to the relationship between HPCs and the essences of natural kinds. On this alternative reading, HPCs are not themselves to be identified with the essences of natural kinds. Instead, the HPC correlated with a natural kind constitutes something like its "operational definition" that can be used to pick out some other property that is the kind's genuine *a posteriori* essence (or "natural definition"). While there are some passages in Boyd (1988) that *prima facie* permit this alternative interpretation, other passages—especially in Boyd's later writings—strongly favor my preferred interpretation according to which the HPC *just is* the essence or natural definition of the relevant natural kind. For instance, Boyd writes: "I conclude that individual species have (homeostatic property cluster) essences, so that a form of 'essentialism' is true for species…" (1999a: 142); "…there are a number of scientifically important kinds…, biological species among them, whose natural definitions are very much like the property-cluster definitions postulated by ordinary-language philosophers except that the unity of the properties in the defining cluster is mainly causal rather than conceptual" (1999c: 67; cf. 1988: 196); "Species are defined, according to the HPC conception, by those shared [phenotypic] properties *and* by the mechanisms…which sustain their homeostasis" (1999b: 81, emphasis in the original). For additional passages that favor my interpretation see note 12 below.

[5] The notion that natural kinds are defined by clusters of properties can be found in Mill (1867) and Russell (1948). Russell's own account strikingly anticipates Boyd's. Russell writes: "The essence of a natural kind is that it is a class of objects all of which possess a number of properties that are not known to be logically interconnected" (1948: 317). His claim that the defining properties are not known to be logically interconnected suggests that he recognizes that their belonging to the kind's essence is not a matter of our linguistic conventions but rather is a matter of a nomological connection. Furthermore, Russell backs away from the claim that every member of a kind needs to share all of the kind-defining properties. Like Boyd, he accepts indeterminacy in the extensions of natural kind terms: "Assuming evolution, there must have been outlying members so aberrant that we should hardly know whether to regard them as part of the [intension] or not" (*ibid*., 443).

nearly every) member of the same species, conspecifics are uniform with respect to very many manifest properties. As a result of this uniformity, when we observe that all observed samples of a species exhibit a property G, we can often reliably infer that all unobserved samples will exhibit G as well. This remains true even when our original sample is relatively small (Boyd 1999b: 82; Kornblith 1993: 92ff; Russell 1948: 318). For example, biologists were no doubt able to infer (with a high degree of epistemic warrant) that all female platypuses are egg layers upon observing only a few specimens that laid eggs.

Traditionally, the essence of a natural kind is thought of as a property or a collection of properties whose exemplification by a given individual is both necessary and sufficient for that individual to count as a member of the kind. This sort of view is reflected in the claim that necessarily something is a quantity of WATER if and only if it has the property *being* H_2O, where *being* H_2O is understood to be the essence of WATER. Boyd relaxes this requirement so that an individual may belong to an HPC kind even if it fails to instantiate some of the properties in the kind's HPC essence.[6] He writes: "Imperfect homeostasis is nomologically possible or actual: some thing may display some but not all of the properties in [the property cluster] F; some but not all of the relevant underlying homeostatic mechanisms may be present" (1988: 197; 1999a: 143). As long as the individual instantiates "enough" of the "important" properties in F, where these properties are unified by "enough" of the relevant mechanisms, it is properly classified as a member of the kind whose essence is constituted by F and F's homeostatic

[6] Boyd's HPC conception of natural kind real essences also departs from more traditional views insofar as it denies that the essences of natural kinds must be (i) "unchanging" and (ii) composed only of properties that are both "ahistorical" and (iii) intrinsic to the kind's members (Boyd 1999a: 146f, 153-157; cf. Ellis 2001: 19-23).

mechanisms. This feature of Boyd's view allows us to class certain anomalies, such as mutants, within the species to which they intuitively belong. Moreover, it permits cases of indeterminacy where there are

> …things that display some but not all of the properties in F (and/or in which some but not all of the relevant homeostatic mechanisms operate) such that no rational considerations dictate whether or not they are to be classed under [natural kind term] t, assuming that a dichotomous choice is to be made (1988: 197).

If biological species are to count as HPC kinds, then this relaxed understanding of natural kind essences must be accepted. Because of the gradual nature of evolution, there are bound to be cases in which it is indeterminate whether some particular individual is a member of a given species.[7]

5.2.3. Two Examples of HPC Kinds.

Boyd offers biological species as paradigmatic examples of HPC kinds. He also suggests that certain chemical kinds are examples of HPC kinds. To illustrate the HPC conception of kinds, I will consider an example from each of these domains. Although Boyd is most emphatic that biological species are HPC kinds, I find it helpful to begin with an example from the domain of chemistry.

Boyd seems to accept that chemical elements such as GOLD and compounds such as WATER are natural kinds with real essences that conform to the more traditional (non-HPC) conception of essences: their essences identify fully necessary and sufficient

[7] Not only is the possibility of imperfect homeostasis and extensional indeterminacy important for the plausibility of HPC definitions of biological species, it also plays a role in Boyd's defense of moral realism. Some have thought that the existence of actions whose moral status is irresolvable is best explained by the hypothesis that there are no moral facts (see, e.g., Mackie 1977: 37). Boyd's HPC account of moral properties makes an alternative explanation possible. If moral terms designate HPC phenomena, then, as with species, we should expect instances where it is indeterminate whether or not an individual action or state of affairs falls within the extension of a given moral term. Thus, not only are such indeterminate cases not an embarrassment to an HPC conception of moral properties, they are predicted by it (Boyd 1988: 213).

conditions for membership in these respective kinds. Still, he suggests that other, more general, chemical classifications may mark HPC kinds. One example of a chemical HPC kind that Boyd offers is METAL (1999b: 83f). He proposes that the cluster of properties that define METAL includes (among other properties) *conductivity*, *ductility*, *malleability*, and the property of *having an inverse relationship between conductivity and temperature*. These properties (and others) are regularly co-instantiated in distinct individual quantities of substance. Boyd does not say what things serve as the homeostatic mechanisms that unite these properties in samples of METAL. A plausible candidate for such a mechanism is something like the property of *being composed of atoms that donate electrons*.[8] In virtue of their homeostatic relationship, the aforementioned collection of properties satisfies PH and thus constitutes an HPC. Boyd's suggestion is that this HPC, along with its homeostatic mechanisms, constitutes the *a posteriori* real essence of the kind METAL: a portion of substance is a piece of METAL when and only when it instantiates enough of these properties (weighted for importance) where their co-instantiation is due (at least in part) to mechanisms such as (e.g.) *being composed of atoms that donate electrons*.

Consider next the biological species TIGER. Any individual tiger instantiates innumerably many common morphological, physiological and behavioral properties. Among these are properties corresponding to its particular skeletal structure, the arrangements of its organs, its behavioral dispositions etc. For an individual tiger, its particular genotype is a plausible candidate for the (most central) underlying mechanism that causes and sustains the co-instantiation of its (intrinsic) properties. However, when

[8] Here and elsewhere I treat the relevant homeostatic mechanisms as properties. I do so in this case because the mechanism associated with the kind METAL is evidently something that has multiple instances. Although Boyd is not explicit about what the ontological status of the homeostatic mechanisms is supposed to be, his own examples are also of things that admit of multiple instantiations. In any case, nothing much here turns on whether we understand homeostatic mechanisms as properties rather than individuals.

we turn our attention to the biological species TIGER itself (as opposed to its individual members), discerning the defining homeostatic mechanisms is somewhat trickier than it is for chemical kinds. We might be tempted to suppose that the TIGER genotype is the sole homeostatic mechanism in the definition of the kind TIGER; but this is not Boyd's view. If I understand him correctly, one reason for rejecting such a view is this: without additional mechanisms, such as "*gene exchange between certain populations* and *reproductive isolation from others*,"[9] the properties that define TIGER might fall out of homeostasis in a relatively short period of time. For instance, if tigers were not reproductively isolated from other biological species, they might interbreed with them and bring new genes into the TIGER gene pool. In turn, some of the manifest properties found in the defining cluster may be quickly lost. For example, tigers might lose their stripes or their tails if a new dominant gene were to spread throughout their population. What this illustration shows is that some of the mechanisms responsible for the homeostatic unity of the properties that define the kind TIGER are extrinsic and external to individual tigers.[10] These mechanisms (both internal and external ones), along with the morphological, physiological, and behavioral properties that are produced, maintained, and unified by them, constitute the HPC essence of the kind TIGER (cf. 1999a: 142; 1999b: 81). An individual is a tiger just in case it instantiates enough of these properties

[9] Boyd offers these and other examples of homeostatic mechanisms unifying the properties of biological species in his (1999a: 165; cf. Mayr 1996). I have added italics in keeping with my convention of italicizing property-referring terms.

[10] Of course, given the evolution of biological species, the homeostatic unity of certain property clusters is bound to be disturbed over a long enough period of time. This implies that the constituents of a biological kind's HPC definition change over time. Boyd recognizes and accepts this consequence of his view. He writes, "…the properties which determine the explanatory definition of a species (and, thus, the conditions for membership in it) may vary over time (or space), while it continues to have numerically the same definition" (1999c: 68).

(weighted for importance) where their co-instantiation is due (at least in part) to the above-mentioned mechanisms.

5.3. Homeostatic Consequentialism: THE MORAL GOOD as an HPC Kind.

Boyd proposes that THE MORAL GOOD, like METAL and THE TIGER, is an HPC kind. On this view, the property *moral goodness*[11] is itself constituted by a cluster of properties that are homeostatically unified.[12] The properties that compose *goodness* correspond to "things which satisfy important human needs" (Boyd 1988: 203). Here, Boyd gestures toward what those needs are:

> Some of these needs are physical or medical. Others are psychological and social; these (probably) include the need for love and friendship, the need to engage in cooperative efforts, the need to exercise control over one's own life, the need for intellectual and artistic appreciation and expression, the need for physical recreation, etc. (*ibid*.).

Boyd does not say which properties in fact correspond to the satisfactions of these needs. The following seems like a plausible (though perhaps not exhaustive) list of properties whose instances are the satisfactions of the human needs Boyd adumbrates above: *being educated*, *being physically healthy*, *sharing friendship*, *sharing love*, *enjoying leisure*, *engaging in physical recreation*, *engaging in cooperative efforts*, *creating and appreciating art*, and *being autonomous*.[13] As I understand Boyd, these properties (along

---

[11] From here on, I will typically drop the adjective 'moral' from 'moral goodness' and 'the moral good.' Unless otherwise indicated, 'goodness' and 'the good' should be taken to refer to *moral goodness* and THE MORAL GOOD.

[12] As evidence that Boyd means to identify *moral goodness* with an HPC of the sort that I am about to introduce, note that Boyd explicitly writes "…the term 'good' in its moral uses refers to the homeostatic cluster property…" (1988: 205). Note also that my reading of Boyd as identifying *goodness* with an HPC also accords with the way Nicholas Sturgeon—himself a supporter of the HPC conception of *goodness*—understands Boyd. He writes that Boyd thinks of "…moral properties such as intrinsic goodness as homeostatic clusters of various natural features…" (Sturgeon 2003: 550).

[13] It should be clear from this list that I understand the properties that putatively compose *goodness* to be properties whose instances satisfy the various human needs (where, for example, instances of *being in love*

with certain homeostatic mechanisms to be mentioned shortly) are constitutive of *goodness*. They are also constitutive of the essence and definition of THE GOOD. Thus, we should think of the property *being physically healthy* as playing a role in the HPC that defines THE GOOD that is analogous to the role that (e.g.) *malleability* plays in the HPC that defines METAL and analogous to the role that (e.g.) *being quadrupedal* plays in the HPC that defines THE TIGER. Following Boyd, I will refer to the properties that are constitutive of *goodness* as 'the human goods' (*ibid*.).

If the human goods are to constitute an HPC essence, they must be homeostatically clustered. That is to say, they must satisfy one of the three disjuncts on the right-hand side of PH. Boyd's view seems to be that they satisfy the last disjunct of PH (i.e., the conjunction of clause i and clause ii):

> Under a wide variety of (actual and possible) circumstances these human goods (or rather instances of the satisfaction of them) are homeostatically clustered. In part they are clustered because these goods themselves are—when present in balance or moderation—mutually supporting. There are in addition psychological and social mechanisms which when, and to the extent to which, they are present contribute to the homeostasis. They probably include cultivated attitudes of mutual respect, political democracy, egalitarian social relations, various rituals, customs, and rules of courtesy, ready access to education and information, etc. It is a complex and difficult question in psychology and social theory just what these mechanisms are and how they work (1988: 203).

If Boyd is right, then the human goods can be said to constitute an HPC. It is natural to suppose that this means that something like the following is true: under suitable social and psychological conditions, whenever an individual person is (e.g.) happy and enjoys leisure, an education, autonomy, and cooperative efforts, there is an increased likelihood that that same individual is also physically healthy, engages in physical recreation, shares

---

satisfy the need for love). One might be tempted to read Boyd as claiming instead that the properties that compose *goodness* are instances a broader property, *viz*., *having a need satisfied*. I do not think such a reading could be correct. If it were, then there would be a single property that composes *goodness*: *viz*., the property *having a need satisfied*. In that case, *goodness* could not be thought of as a *cluster* of properties.

friendship and love, and appreciates art. Boyd, however, is anxious to point out that the homeostasis between instances of the human goods need not involve their all being possessed by one and the same individual. He writes: "The properties in homeostasis are to be thought of as instances of the satisfaction of particular human needs among people generally, rather than [merely] within the life of a single individual" (1988: 204n). His idea seems to be that, when the relevant homeostatic mechanisms (e.g., democracy, social equality, etc.) are in place, there is a causally sustained tendency for the having of some goods by one or more individuals to bring about or sustain the having of these and other goods by other individuals as well. For example, under the proper social conditions, Bob's engaging in artistic activity and physical recreation (etc.) will bring about, sustain, or otherwise enhance Carol's appreciation for art, her education, and her own engagement in physical recreation (etc.).

Boyd's central claim is that (something like)[14] this HPC and the mechanisms that unify it define THE MORAL GOOD:

> [THE MORAL GOOD][15] is defined by this cluster of goods and the homeostatic mechanisms which unify them. Actions, policies, character traits, etc. are morally good to the extent to which they tend to foster the realization of these goods or to develop and sustain the homeostatic mechanisms upon which their unity depends (1988: 203).

[14] It should be acknowledged that Boyd presents his particular account of the human goods and the homeostatic mechanisms that unify them as speculation. He is careful to note that the question of exactly which properties and mechanisms belong to the cluster that defines THE GOOD is a matter for empirical inquiry. The success of the HPC conception of *goodness* does not depend upon the correctness of precisely this list of goods and mechanisms (although Boyd believes that his characterization of the HPC is "close to the truth" [1988: 202]). With the exception of the argument I offer in §5.5.2, my arguments against Boyd's view can be directed against other HPC proposals of *goodness* with little or no modification.

[15] In the original text, Boyd uses 'moral goodness' where I use 'the moral good.' I have modified this passage in order to preserve the symmetry between moral HPC kinds and biological and chemical HPC kinds. Thus, although he writes that *moral goodness* is defined by the cluster of goods and their homeostatic mechanisms, I take him to mean that the property *moral goodness* is *constituted* by this cluster and its mechanisms. In turn, the HPC *moral goodness* defines the kind THE MORAL GOOD.

Although what Boyd proposes here is a theory of value and not a theory of right action, I will follow him in calling this account of THE GOOD *homeostatic consequentialism*.[16]

If THE GOOD is defined by the cluster of human goods and its homeostatic mechanisms, then it is natural to suppose that something (e.g. a state of affairs) is good—i.e., is an instance of *goodness*—just in case it instantiates the cluster of human goods (where these goods are unified by the relevant mechanisms). Let's call this proposal "HC1":

> HC1: A state of affairs, P, is non-instrumentally[17] morally good if and only if P instantiates the HPC of human goods.

The second sentence in the last passage quoted might be thought to be in tension with HC1. That sentence suggests that an entity can be good even if it does not itself instantiate the cluster. All that is needed is that the entity "tends to foster the realization" of the cluster. This would be a very surprising possibility if THE GOOD were an HPC kind. After all, we do not say that some policy that tends to foster the realization of *tigerhood* (e.g., a policy of breeding tigers or protecting them from hunting) is itself a tiger (i.e., an instance of *tigerhood*). Only those entities that instantiate the cluster of

---

[16] Boyd evidently thinks of homeostatic consequentialism proper as a broader moral theory that includes the present HPC account of *goodness* as just one component. This larger view would presumably include a consequentialist account of *moral obligation* alongside the HPC conception of value. (Boyd discusses his consequentialist view of right action in greater depth in his [2003b: 24-47].) However, since he does not offer a distinct name for the theory of value he proposes, it will be convenient for our purposes to use 'homeostatic consequentialism' to denote only the HPC theory of value.

[17] Just below, I explain why making a distinction between *non-instrumental goodness* and *instrumental goodness* is desirable for the homeostatic consequentialist. Although Boyd does not himself acknowledge the distinction in his (1988), fellow homeostatic consequentialist Nicholas Sturgeon attributes to Boyd the view that homeostatic consequentialism is an account of *intrinsic goodness* (Sturgeon 2003: 550). However, Sturgeon suggests that the kind of *intrinsic goodness* he has in mind is not "a property that depends only on the intrinsic, nonrelational properties of the things that have it" (*ibid*.). Because *intrinsic goodness* is sometimes thought to be just the sort of *goodness* that a thing has in virtue of its intrinsic, non-relational properties, I prefer to label the sort of *goodness* that Sturgeon describes as "non-instrumental." I cannot here attempt a precise account of the distinction between *instrumental* and *non-instrumental goodness*. Perhaps a slogan will suffice: "a thing is non-instrumentally good if and only if it is good as an end, rather than good merely as a means to some other good thing." For a useful discussion, see Kagan (1998). (Note, however, that Kagan takes the label 'intrinsic goodness' to apply both to *non-instrumental goodness* and to the kind of *goodness* a thing has in virtue of its intrinsic, nonrelational properties.)

properties that defines the kind TIGER are tigers. For this reason, I think it is best to read the second sentence in the last quoted passage as describing a phenomenon that is distinct from *moral goodness* or THE MORAL GOOD. We might call this phenomenon "instrumental moral goodness."[18] Roughly, something is instrumentally morally good on this view just in case it tends to foster the realization of the HPC that defines THE GOOD. My focus will be on homeostatic consequentialism as an account of *non-instrumental moral goodness*.

5.4. The Case against Homeostatic Consequentialism.

5.4.1. Isolated goods.

Although homeostatic consequentialism is presented foremost as a metaethical view, it has substantive moral implications. Consider the following scenario. There is a hermit, alone in the woods. Although it is a relatively cool day, the sun peeks out from behind the clouds and warms the hermit's back. The hermit finds this sensation pleasurable. Suppose, however, that this pleasure contributes neither to his nor anyone else's having friends. Nor does it contribute to his appreciation of art, his engagement in cooperative efforts, his sharing love, etc. In short, the hermit's experience of pleasure causally

[18] Here are two additional considerations in support of this exegetical decision: First, as we saw earlier, Boyd represents himself as claiming that "…the term 'good' in its moral uses refers to the homeostatic cluster property just described…" (1988: 205). This way of representing homeostatic consequentialism is hard to square with the view that an act is good just in case it "tends to foster" the HPC of human goods. For if that were Boyd's view, he should have said that 'good' (or better, 'goodness') refers to the property of *tending to foster the HPC just described*. A second consideration concerns the sorts of items that Boyd cites as bearers of *moral goodness* in the passage cited above. The three kinds of items he cites are actions, policies, and character traits. However, on a standard consequentialist conception of morality, only states of affairs are taken to be the fundamental bearers of non-instrumental (or intrinsic) value; actions, policies, and character traits are typically understood to have only instrumental (or extrinsic) value. This gives us yet more reason to treat the passage cited above as describing *instrumental moral goodness*, a property that is distinct from the more fundamental *non-instrumental moral goodness*.

contributes to the realization of very few, if any, of the human goods.[19] It follows that the hermit's being pleased does not instantiate the HPC that putatively constitutes *goodness*. Even if we thought that the property *being pleased* is itself a human good, the instantiation of only one property in a cluster is not the same thing as the instantiation of the cluster itself.[20] Because the hermit's being pleased fails to instantiate the HPC that putatively constitutes *goodness* (and, moreover, fails to contribute to the realization of that HPC), HC1 implies that the hermit's experience of pleasure is not good. It follows that the world in which the hermit experiences this particular episode of pleasure is no better than a world that is otherwise identical except that the hermit does not experience this pleasure. This consequence of HC1 is surely counterintuitive. The world in which the hermit experiences the additional pleasure is the better world. If so, we must reject HC1. Let us call this objection to homeostatic consequentialism *the problem of isolated goods*.

(While I have taken the property *being pleased* as my example of an isolated good, it should not be thought that the objection depends upon this choice. For those who are not inclined to view *being pleased* as a good-making property, we can modify the hermit example so that the relevant state of affairs realizes some other putative good-making property in causal isolation from the rest of the human goods. For example, we might imagine instead a state of affairs in which the hermit has a preference satisfied—or appreciates the beauty of some landscape or contemplates some magnificent truth, etc.—

[19] There is probably some correlation between experiences of pleasure and a person's physical health. So we may have to grant that this particular episode of pleasure makes a causal contribution to the hermit's health, however slight.

[20] In light of the discussion of §5.2.2, it should be observed that an HPC "as a whole" may be instantiated by an individual even when some of the cluster's constituent properties are not instantiated by that individual. Even so, it should also be clear that an HPC itself is not instantiated in an individual that instantiates only one of its constituent properties, e.g. an individual's being striped is not sufficient for it to be properly classified as a TIGER.

where this state of affairs fails to causally contribute to the realization of other human goods.)

5.4.2. An alternative formulation.

Hitherto, I have taken HC1 to express the core thesis of homeostatic consequentialism. However, there may be a different way to understand the view:

> HC2: A state of affairs, P, is non-instrumentally morally good if and only if P instantiates at least one of the human goods in the HPC of human goods.

One benefit of HC2 is that, by allowing there to be a plurality of non-instrumental good-making properties, it brings homeostatic consequentialism closer to more traditional forms of axiological pluralism. More important, however, it promises an answer to the problem of isolated goods. Here is how. Suppose that *being pleased* is a human good. Although the state of affairs in which the hermit is pleased does not instantiate the entire cluster of human goods, it does instantiate at least one human good that is part of the cluster. Given HC2, this is sufficient for it to be true that the hermit's being pleased is non-instrumentally good.

Unfortunately, the problem of isolated goods returns in a slightly different form to threaten even HC2. Consider a possible world, W, in which the homeostatic mechanisms that putatively unify the human goods in our world are absent. In W, the socio-political conditions are such that it is not true that the presence of some of the human goods raises the likelihood that the others will be present. (We might imagine that the human social environment in W is something like a Hobbesian state of nature.) From this assumption, it follows that the human goods are not homeostatically clustered in W. Next, imagine that the sunbathing hermit is in W. Once again, his pleasure neither instantiates nor

contributes to the instantiation of a larger cluster of human goods. More important, however, because the hermit is in W, his pleasure does not instantiate a property that is homeostatically clustered with other human goods. From this, it follows that, even on HC2, this hermit's pleasure is not good. As before, this consequence of HC2 clashes with considered moral judgment.[21] (And again, the objection could be restated *mutatis mutandis* with some other putative good-making property in place of *being pleased*.)

At this point, the homeostatic consequentialist might appeal to something like the idea of rigid designation in order to answer the "revived" problem of isolated goods. The strategy is roughly this: Let us grant that there is a homeostatically unified cluster of human goods here in the actual world. The predicate 'non-instrumentally good' applies to any instance of one (or more) of those human goods. This predicate should be understood to apply "rigidly." A predicate is a rigid applier, according to this strategy, just in case, if it applies to all instances of a property F in the actual world, it applies to all instances of F in every possible world.[22] Since 'non-instrumentally good' presumably

---

[21] There is some indication that a homeostatic consequentialist would be willing to bite the bullet here. Sturgeon allows that "nothing would have a property such as intrinsic goodness at all, given a radical enough breakdown" in a certain part of the HPC that constitutes *goodness* (Sturgeon 2003: 550). Since the Hobbesian world exhibits a breakdown of the HPC that constitutes *goodness*, it would appear that Sturgeon is willing to accept that the hermit's being pleased is not an instance of *goodness*. (On the other hand, Sturgeon is here speaking of a breakdown only in a specific part of the HPC. In particular, he is considering the breakdown in the part of the cluster that involves "the existence of purposive, valuing creatures somewhat like us" [*ibid*.]. The claim that a world without purposive, valuing creatures contains no *goodness* strikes me as far less controversial than the claim that a world that included such creatures would nevertheless fail to contain intrinsic or non-instrumental value, if there were a lack of homeostasis between the human goods.)

[22] I do not claim that this is the best or most useful conception of rigid application. It is, however, the conception that the defender of HC2 needs in order to avoid the revised problem of isolated goods. As will be seen below, I think this conception of rigid application is defective. An arguably better conception can be found in Devitt (2005). Unfortunately, Devitt's conception is of no help to HC2. Furthermore, I doubt that the more traditional notion of rigid designation could be deployed in the service of HC2 without making questionable assumptions. On the traditional view, a term, t, rigidly designates an entity, e, if and only if t designates e in every possible world in which e exists, and t designates no other entities (Kripke 1971: 78, 79). As an initial difficulty, it isn't clear that predicates are the sort of items that can be rigid designators: if predicates (be they natural kind predicates or nominal kind predicates) designate their extensions, then none are rigid, since their extensions are different at different possible worlds. If they

applies to all instances of *being pleased* in the actual world, as a rigid applier it also applies all instances of *being pleased* in every possible world. As a result, it is consistent with HC2 to ascribe *non-instrumental goodness* to the hermit's being pleased in the Hobbesian world. This is so despite the fact that the human goods are not homeostatically clustered in that world.

It is doubtful that this conception of rigid application is defensible, at least if it is supposed to capture some semantic property common to natural kind terms. Suppose that 'cordate' is a rigidly applying term. It applies to all actual instances of the property of *being a creature with a heart*. As a rigid applier, it also applies to all possible instances of *being a creature with a heart*. The trouble begins when we notice that 'cordate' also applies to all actual instances of the property *being a creature with a kidney*. From this, along with the assumption that 'cordate' is a rigid applier, it follows that 'cordate' applies to all possible instances of *being a creature with a kidney*. But this is false.

I think that this shows that this conception of rigid application is defective. If I am right, then the homeostatic consequentialist will not be able to appeal to it in his

---

designate properties, then every meaningful predicate is a rigid designator. In that case, rigid designation marks no interesting distinction among predicates. This last consequence might be avoided if we suppose that only those predicates that designate sparse properties (or universals) rigidly designate. Assuming this restriction were defensible, it would still require some maneuvering to get this conception of rigid designation to do the work HC2 needs of it. For one thing, given the pluralistic assumptions of HC2, there is no one property that 'good' designates; there is a plurality. Thus, it is not true of 'good' that it designates an entity in every possible world in which that entity exists *and designates no other entities*. One solution is to suppose that 'good' designates the conjunctive property made up of all the different good-making properties. But now HC2 has no reply to the problem of isolated goods. As we saw, the hermit's episode of pleasure does not instantiate a conjunctive property that includes all the other putative good-making properties as constituents. What is needed instead is an account where 'good' designates a disjunctive property. Here we face more trouble. We have had to assume that rigidly designating predicates designate only sparse properties. However, on familiar conceptions of sparse properties, disjunctive properties do not qualify as sparse (Armstrong 1978: 19-23). Perhaps there is more that can be said that would make it plausible that rigid designation can do the work that HC2 needs it to do. I hope to have said enough to make it clear that rigid designation does not provide a quick or easy solution to the problem of isolated goods.

response to the revised problem of isolated goods. However, it might be thought that the argument of the previous paragraph shows only that it was wrong to assume that 'cordate' is a rigid applier. This is a desperate tack. As a biological kind term, 'cordate' is a good candidate for a natural kind term (and possibly even an HPC kind term). But if 'cordate' is a natural kind term, then some explanation is required for why it fails to be a rigid applier whereas a supposed (HPC) natural kind predicate like 'good' succeeds.

5.4.3. Two structural disanalogies.

Even if HC2 could avoid the revived problem of isolated goods, the move from HC1 to HC2 is suspiciously *ad hoc*; there is no precedent for a view like HC2 in the general theory of HPC kinds. Let me explain.

Given HC2, each of the individual properties (i.e., human goods) that compose the HPC that putatively defines THE GOOD is such that it is proper to predicate *non-instrumental moral goodness* of its instances. For example, it is proper to say of John's being in love with Mary that it is morally good. Likewise, we can say that Sam's creating and appreciating art at some particular time is morally good. Moreover, we can say that Rachel's being healthy is morally good.

No paradigmatic HPC kind is like THE GOOD in this respect. Consider the properties that define TIGER. We do not say of a particular tiger that its being quadrupedal is a tiger. Nor do we say that its stalking behavior is a tiger. Nor do we say that its being warm blooded is a tiger. In general, the property of *being a tiger* does not belong to the instances of the individual properties that define TIGER.[23] The same

[23] Of course an instance of a property like *being quadrupedal* might be a part of a larger state of affairs that constitutes some individual's being a tiger. But this does not mean that the state of affairs consisting in a

observation holds for other biological kinds as well as for chemical kinds (e.g., the *malleability* of this piece of metal is not itself an instance of METAL).[24]

What this shows is that, as it is characterized by HC2, *goodness* is structurally unlike the property clusters that define paradigm HPC kinds. *Goodness* is a property exemplified by instances of the individual properties that (putatively) constitute it. Paradigmatic natural kinds like species and chemical substances do not share this feature. This ought to make us suspicious of HC2. If *goodness* really were constituted by an HPC, we would expect it to have the same metaphysical structure as the HPCs that define paradigmatic HPC kinds. Unless there is a convincing precedent for an HPC that functions as *goodness* does according to HC2, the move to HC2 would appear to be *ad hoc*.[25] As far as I can see, the only motivation for HC2 is its promise to solve the first version of the problem of isolated goods.

Since both HC1 and HC2 are vulnerable to the problem of isolated goods, this most recent objection to HC2 would seem to make HC1 the more attractive statement of homeostatic consequentialism. Unfortunately, there is another structural feature of paradigmatic HPC kinds that simply cannot be extended to THE GOOD without absurdity. (This feature creates trouble for homeostatic consequentialism on either of its formulations.)

---

given individual being quadrupedal itself has the property of *being a tiger*. In such a case, we should say instead that one and the same individual has the property of *being quadrupedal* and has the property of *being a tiger*.

[24] In §5.5.4 I introduce putative examples of HPC social kinds. It should be noted here that even those kinds behave like the paradigm HPC kinds and not like THE GOOD. For instance, suppose that the property of *keeping kosher* is part of the cluster of properties that defines the social kind HASIDIC JEW. Some particular man's keeping kosher does not have the property of *being a Hasidic Jew*.

[25] Boyd suggests that the predicates 'healthy' and 'is healthier than' express HPC phenomena (1988: 198). It may be that my challenge (to find a paradigmatic example of an HPC that shares the structural features of *goodness* as understood by HC2) could be answered by developing a plausible HPC account of HEALTH. Unfortunately, Boyd says very little about what sorts of properties might compose the HPC that defines HEALTH. In the absence of a more detailed account, it is difficult to tell whether or not the example of HEALTH will help the homeostatic consequentialist meet the present challenge.

For paradigmatic HPC kinds, most (though perhaps not all) of the properties that are part of the kind's definition are properties had by individual members of the kind. For example, just as *ferociousness* and *being quadrupedal* are part of the HPC definition of the kind TIGER, individual tigers are themselves ferocious and quadrupedal. Likewise, just as *malleability* and *conductivity* are part of the definition of METAL, individual pieces of metal are malleable and conductive.

By contrast, the properties that putatively define THE GOOD are not had by individual members of THE GOOD: no good state of affairs is pleased, educated, enjoys leisure, or shares love etc.[26] To predicate one of these properties of a good state of affairs is to commit a category mistake. Once again, we are presented with a way in which the cluster of properties that putatively defines THE GOOD fails to behave like the property clusters that define paradigmatic examples of HPC kinds. This provides yet another reason to doubt that THE GOOD is an HPC kind.

5.4.4. <u>An alternative cluster of properties.</u>

Both of the structural disanalogies just described might be avoided if we take a rather different collection of properties to constitute the cluster that putatively defines THE GOOD. Consider the following list of properties: *being pursued by rational beings*, *being worthy of being loved*, *meriting realization*, *being approved of by ideal observers*, *being fitting*, and *deserving appreciation*. If properties of this sort were taken to compose the

[26] Of course, individual persons that are constituents of these states of affairs might have these properties. However, because states of affairs are the primary basic bearers of *goodness*, this is of little help. The current difficulty might be avoided by noting that persons themselves may reasonably be taken to be bearers of *moral goodness*. This reply helps if we are advancing some sort of HPC account of virtue. Still, I presume that Boyd and other homeostatic consequentialists want to say that things other than persons may be non-instrumentally morally good. If so, they are faced with the present difficulty.

cluster that defines THE GOOD, then the structural dissimilarities between the good and other HPC kinds would be avoided. With respect to the first disanalogy noted above, it is arguably not a serious defect if we could not properly say of a particular state of affairs that the fact that it merits realization is (non-instrumentally) good. With respect to the second disanalogy, we make no category mistake when we say that a particular good state of affairs merits realization, is fitting, or is deserving of appreciation, etc. Perhaps, then, a collection of properties such as this could serve as the cornerstone of a different proposal for an HPC definition of *goodness* (one that would replace Boyd's own proposal).

There are at least two reasons for thinking that this maneuver will not be successful for the homeostatic consequentialist. First, all of the aforementioned properties contain at least one unabashedly normative property as a constituent: *worthiness*, *merit*, *fittingness*, *desert*, *being ideal*, and *being rational*. An HPC definition of *moral goodness* that includes such properties does nothing to advance the naturalistic accommodation of moral properties that Boyd and other naturalist moral realists are pursuing. After all, it is the putative normativity of moral properties that has led so many philosophers to think that such properties cannot be accommodated within a naturalistic metaphysic (Ayer 1952: 105; Blackburn 1984: 187ff; Hare 1952: 91; Mackie 1977: 38ff; Stevenson 1944: 336).[27] Without a further naturalistic account of these normative properties, homeostatic consequentialism will have accomplished very little, if anything, in the way of showing that *goodness* can be admitted into a naturalistic ontology.

[27] In fact, in selecting these particular properties as potentially definitive of THE GOOD, I was inspired by various non-naturalist and constructivist analyses of 'good.'

A second worry is this: to the extent that properties like *being fitting*, *deserving appreciation*, and *being worthy of being loved* (etc.) are clustered in nature, their clustering does not seem to be a matter of causal connection. In the first place, several of these properties may well be identical. It may be, for example, that the property designated by 'meriting realization' is identical to the property designated by 'being fitting.' Second, even where distinct properties are designated by these terms, the connection between these properties appears to be conceptual or metaphysical rather than causal or nomological. It strikes me as a conceptual truth that an ideal observer is one who approves of all and only that which merits realization or deserves appreciation. To my ears, the claim that there is an ideal (moral) observer who, nevertheless, approves of that which does not deserve appreciation sounds incoherent.[28] At any rate, even if the necessity involved is not conceptual, it is doubtful that the laws of nature could have been different in such a way that there exist ideal observers who approve of that which does not deserve appreciation or that which does not merit realization. If so, the necessity binding these properties is stronger than mere nomological necessity; it is better characterized as a metaphysical necessity. My point here is that a property cluster in which these new properties (*being fitting*, *being deserving of appreciation*, etc.) were given significant definitional weight would not fit Boyd's characterization of an HPC. The property cluster under consideration appears to be, in most cases at least, unified by conceptual or metaphysical necessity.[29] By contrast, the necessity that unifies the

[28] A. C. Ewing expresses roughly the same thought. He suggests that the claim that what is good or right is what an impartial spectator would approve of "is equivalent to saying that something is good or right when it is approved by somebody who only approves what is really good or right" (Ewing 1953: 85).

[29] Of the properties I have recommended for this alternative HPC proposal, the one exception seems to be the property of *being pursued by rational beings*. If one takes a Humean or instrumentalist view of rationality, then it will be at best a metaphysically contingent fact that rational beings pursue, e.g., fitting states of affairs. Thus, *being pursued by rational beings* may well be nomologically linked to the other

properties of an HPC is causal or nomological. Thus, even on this alternative proposal, it would not be true that *moral goodness* is an HPC.

5.5. Inductive Inference and THE GOOD.

5.5.1. Outline of the Argument.

In §5.2.1, we saw that Boyd takes natural kinds to provide the metaphysical ground for successful induction and explanation. His HPC conception of natural kinds is meant to account, in part, for how it is that (some) natural kinds play this role. In §5.2.2, I explained that the homeostatic clustering of properties makes the instances of an HPC kind fairly uniform with respect to their manifest properties. In turn, this uniformity makes reliable inductive inference possible. If this is so, and if THE GOOD is an HPC kind, then we should expect that THE GOOD (and *goodness*) will ground significant reliable inductive inferences. To the extent that this expectation is not fulfilled, we have reason to think that THE GOOD is not an HPC kind. In the next two sections I argue that there is, in fact, reason to think that *goodness* does not facilitate significant reliable inductive inferences. If I am right, then we have even more reason to doubt that THE GOOD is an HPC kind. (As far as I can tell, my arguments remain cogent regardless of whether homeostatic consequentialism is understood as HC1 or understood as HC2. Still, it may aid the reader to know that the following sections were written with HC1 in mind.)

---

properties mentioned. I doubt that this one exception can give the homeostatic consequentialist what he needs to get past the present worry. However, if more properties of this sort could be found, and if there were a compelling case to be made that these other properties are indeed contingently clustered, then this worry could be put to rest. But new difficulties are likely to arise. My suspicion is that the sorts of properties that are needed here would consist primarily in various sorts of characteristic human responses to good states of affairs. If I am right about this, then an alternative HPC definition of THE GOOD that incorporated these properties would raise its own problems for Boyd's larger project of defending naturalistic moral realism. Briefly, a cluster definition involving such properties threatens to make *moral goodness* a response-dependent property. Response-dependent accounts of moral properties, however, are at odds with the robust sort of moral realism that Boyd means to defend.

In arguing that *moral goodness* does not ground significant reliable inductive inferences, I will be relying on anecdotal evidence. For my argument to be conclusive, empirical research would be required. In the absence of such research, I must state the conclusion of my argument modestly. Here I aim to show that, pending the needed empirical research, we ought to be pessimistic as to whether homeostatic consequentialism can make good on its empirical commitments.

5.5.2. Biological kinds versus moral kinds, part I.

For comparison, consider the sorts of reliable inductive inferences that are afforded by biological kinds. While walking in the woods, you see something poking up from behind a log. They are two pieces of furry flesh, about four inches in length, standing straight up. You recognize them as nearly morphologically identical with the ears of some rabbits you have seen. Before you move any closer, you already have a pretty good idea of what you will find as you approach: a furry creature, with short front legs and powerful, kangaroo-like hind legs. The creature will also have whiskers and a short fluffy tail. If you get close enough, you will likely see its nose making a "sniffing" motion. If you get too close, it will rapidly scurry away. In addition to all this, you have a rough idea of what you would find if you were to catch it and cut it open. The background knowledge needed to make these inferences could be culled from having seen only a handful of rabbits in the wild (supplemented with one or two observations of mammalian internal anatomy). On the basis of only a few previous observations, you are able to reliably infer an impressive amount of information about a particular individual

by observing just a pair of ears.[30] Notice, too, that, in this case, you are able to infer the presence of a vast number of properties in the individual from observing comparatively few.

Now consider a putatively moral case. Suppose that, while walking through the park, you observe four young persons playing a game of two-on-two basketball. You observe that each pair is engaged in a cooperative effort and that they are enjoying physical recreation as well as leisure. Each individual appears to be friends with his or her own teammate and, furthermore, all appear to be in good health. Let us grant that all these observations are in fact true. On the basis of these observations, which reveal the instantiation of five out of the nine human goods sketched above (in §5.3), what else can you reliably infer? Well, very likely someone is happy or pleased, as basketball is an enjoyable game. As for some of the other human goods in the proposed cluster, it is anybody's guess. There is no reason to think the game makes any contribution to anyone's education, artistic development, or ability to engage in a loving relationship. I doubt that we can even reliably infer whether playing the game has any impact on anyone's personal autonomy (perhaps two of the players had to be nagged into joining the game). It is worth adding that we do not appear to be warranted in inferring that the state of affairs in which these young people are playing basketball is (all things considered) morally good. In fact, given our evidence, it is not too improbable that the young people are doing something that is, all things considered, bad: perhaps they have

[30] Of course, the reliability of this inference requires the support of some contingent features of the observer's environment as well. For instance, there must not be too many things that look like rabbit ears but are neither attached to rabbits nor creatures that share many (but not too many) properties that are characteristic of rabbits. (Keep in mind, however, that in my example what you infer is not that this thing is a rabbit but rather that it has such and such morphology, anatomy, behavior, etc. If the creature should turn out to be a hare, these conclusions are every bit as correct as if it turns out to be a rabbit).

all neglected their university studies in order to play; perhaps they have unfairly excluded others from joining their game; perhaps one team is in the process of hustling the other team out of their paychecks. If the basketball game exhibits any of these features, then it may well generate enough harm that, on balance, it contributes negative non-instrumental value to the world. What these considerations suggest is that, in fact, the observation of the properties in the cluster that putatively defines THE GOOD does not afford us much, if any, inductive knowledge. In this respect, THE GOOD is nothing like a paradigmatic HPC kind such as THE RABBIT.[31]

Now, if the example just offered is to support the conclusion I want to draw, then it must be generalizable. It will be of no use if I have simply called attention to one of the infrequent cases where some of the properties defining THE GOOD are present but the rest are not. (After all, we sometimes witness small, furry, ear-like things, and they turn out not to be attached to small timid mammals.) What I need to show, then, is that cases in which a number of human goods are present but the others are not constitute the norm rather than the exception. Since I cannot here continue to produce examples of this sort, I will offer a pair of cases of a somewhat different sort to compare. I believe the implications of the following cases are generalizable. They differ from the first pair in that, here, the epistemic agent does not see the individual but is merely told that a given unseen individual belongs to a certain kind. I then consider what sort of information he or she can reliably infer from this (let us grant) accurate testimony about the individual.

[31] To make matters worse, even if we could inductively infer the presence of some of these goods from others, it is not clear that it is reference to THE MORAL GOOD that facilitates these inferences. Our inductions may well turn out to be grounded by the properties that cluster around sporting activities qua sporting activities.

5.5.3. Biological kinds versus moral kinds, part II.

If a person is competent at recognizing members of a natural kind, then she should be able to reliably infer the presence of many properties had by an unseen individual solely by being (truthfully) told that the individual is a member the kind in question. For example, I have seen only a handful of scorpions (real and images thereof); I have also learned several facts about them through the testimony of experts. This relatively small number of encounters has been enough for me to acquire competence in recognizing members of the kind SCORPION. Simply by being told that Snippy is a scorpion, I can reliably infer that Snippy is an insect-like creature, three or four inches long from head to the start of the tail. His tail is roughly the same length as his body and is equipped with a stinger and a pouch full of venom. Moreover, affixed to the front of his body are "lobster-like" claws. In addition to these morphological features, I can reliably infer some behavioral properties: for example, Snippy would eat an insect if it were available; if a person were to agitate Snippy properly, Snippy would sting her. I know all this about Snippy only by being told that he is a member of the kind SCORPION.

Contrast the scorpion case with a case in which we are truthfully informed that a state of affairs, P, that has just taken place is non-instrumentally morally good. We are given no further information concerning P's characteristics. Given our background knowledge of non-instrumentally good things, what inductive inferences would we be justified in making? One might be tempted to infer that P exemplifies whatever property one's favorite axiological theory entails is non-instrumentally good-making. Perhaps, then, P involves the satisfaction of a preference, or someone's being pleased, or both. On the face of it, this hardly seems like an example of inductive inference. Still, if the

homeostatic consequentialist is granted his favored account of moral theorizing, it may well count as a case of inductive inference. Let us suppose, then, that these two inferences concerning P are examples of epistemically justified inductions.

If THE GOOD is an HPC kind, we should expect that we can reliably infer more than this. Unfortunately, I doubt that we can. Consider the human goods listed in §5.3. Surely, the information we have been given about P does not justify us to infer that P involves someone's being in love, engaging in cooperative efforts, or appreciating art, etc. Even a convinced axiological pluralist must recognize that the mere knowledge that P is good does not justify us in inferring which of the good-making properties P realizes. Even less does our information justify us to infer that all (or nearly all) of these human goods are realized in P. This is bad news. If the human goods really are homeostatically clustered and constitutive of *non-instrumental goodness*, we should expect this stronger inference (that nearly all the human goods are realized in P) to be justified. Perhaps there is empirical research that can be conducted that would show that such inferences are reliable. In the absence of this research, however, we have no reason to believe that such a strong inference is epistemically justified. If this is right, then we ought to be skeptical of the claim that *goodness* is constituted by an HPC.

So far, I have been considering the relationship between *non-instrumental moral goodness* and inductive inference. It might be thought, however, that *instrumental moral goodness* is more promising as a ground of reliable inductive inference. Let's consider, then, what we may justifiably infer from the news that Jane has just performed an instrumentally morally good action. I am inclined to think that matters are not much different than before. I suspect that we are justified in inferring that someone was

pleased and had a preference satisfied as a result of Jane's act. I suppose we would be also be justified in inferring that, if anyone's preferences were frustrated as a result of Jane's act (or if anyone was pained by it), this frustration was outweighed by the preferences satisfied (or the quantity of pleasure it brought about). Once again, however, I doubt that we would be justified in inferring that her act causally contributed to the realization of *love*, *friendship*, *cooperation*, *physical health*, and *artistic appreciation*, etc. As often as instrumentally good actions contribute to the realization of these properties, they contribute to their frustration. And again, even if we are in a position to infer that Jane's act contributes to the realization of some of these human goods, we are not in a position to infer which of these it contributes to. Still less are we justified in inferring that it contributes to the realization of nearly all of them.

We might be encouraged to take a long-term view of Jane's action. While her good action might have immediately involved breaking a friendship, ending a love affair, sacrificing someone's autonomy, etc., it may be that, in the long run, her act will contribute to the realization of all these things. If so, then her act does contribute to the cluster of human goods after all. But even if all this turns out to be true of Jane's act, we are surely in no position now to infer this with any kind of confidence. Such an inference may signal an admirable sort of optimism, but it is surely not an example of justified reliable inductive inference. If I am right, then even *non-instrumental goodness* is of little value in grounding significant reliable inductive inference.

To summarize, it is doubtful that the human goods sketched in §5.3 enjoy the sort of homeostatic relationship shared by the properties that define chemical and biological kinds. If the human goods did share such a nomological bond, we would expect to be

able to make a significant number of reliable inductive inferences upon observing that some (at least half) of the properties in the cluster are instantiated by some state of affairs. Furthermore, we would expect that the knowledge that a given state of affairs is good will permit us to reliably infer a significant number of further facts about it. I have argued that neither of these expectations are met. Reference to THE GOOD does not facilitate significant reliable inductions. In this respect, THE GOOD is very much unlike paradigmatic HPC kinds. Consequently, we have yet more reason to doubt that THE GOOD is itself an HPC kind.

It is worth recalling that one of the appeals of Boyd's HPC account of natural kinds is that it purports to explain how such kinds are able to fulfill the inductive and explanatory roles they are alleged to play. In light of this, these observations concerning THE GOOD are no small blow to homeostatic consequentialism. THE GOOD lacks the very feature of natural kinds that the HPC view is meant to account for.

5.5.4. Social kinds.

It might be objected that it is much too demanding to ask that moral kinds ground as numerous and reliable inductions as chemical and biological kinds do. This is a reasonable objection. A fairer comparison might contrast moral kinds with kinds that make up the subject matter of social sciences like psychology or sociology. I take examples of social kinds to include THE STATE, RELIGION, NATIVE AMERICAN, JEW, PSYCHOPATH, HOMOSEXUAL, FOREIGNER, BACHELOR and ECONOMIC DEPRESSION.[32] Boyd suggests that at least some social kinds are HPC kinds.[33] It is reasonable to suppose that,

[32] The first five examples are culled from Richard Miller (2000).
[33] He offers CAPITALISM as an example of a HPC social kind in Boyd (1999b: 83).

if there are HPC social kinds, such kinds support significantly fewer and less reliable inductive inferences than biological and chemical kinds support. Consequently, if social kind terms can be shown to designate HPC kinds, and if reference to THE GOOD facilitates nearly as many reliable inductive inferences as does reference to social kinds, then the arguments of §5.5.2 and §5.5.3 could be answered.

The first challenge facing an objection along these lines is to establish that there are, in fact, plausible examples of social kinds whose extensions are defined by HPCs. At least some of the kinds listed in the previous paragraph seem to resist *a posteriori* HPC definitions. BACHELOR, for instance, has a fairly straightforward analytic definition; even if there turns out to be some properties that contingently cluster around bachelors (*qua* bachelors)—e.g., *having an active night life*—it is doubtful that such properties are part of the definition of BACHELOR. Still, I think HPC social kinds may well exist. At any rate, it seems to me that there exist social kinds that ground interesting reliable inductions where it can be plausibly maintained that the essential properties of these kinds are discoverable only *a posteriori*. The trouble is that these social kinds seem to license a far greater number of reliable inductive inferences than THE GOOD does. In the previous section, I counted only two inductive inferences that seemed to be licensed by the proposition that some particular state of affairs (or action) is good. Contrast this with the inductive inferences afforded by the proposition that some particular individual is a member of the social kind HASIDIC JEW (HASIDUM). Upon being told that Jacob is an adult male Hasidic Jew, we (or at least, those of us somewhat familiar with HASIDUM) can reliably infer that he has a beard and payas (long curls of hair growing from the temples). His typical attire includes a black hat, a black suit with a white button down

shirt, and, on certain occasions, tallis (a prayer shawl) worn under his coat. We can also reliably infer several things about Jacob's weekly activities: he keeps kosher, does not work or drive on the Sabbath, and regularly studies the Torah. Likewise, we can infer that Jacob believes (or at least purports to believe) that the Torah gives a literally true account of historical events.[34]

What this example shows is that, by observing the practices of only several Hasidic Jews, an epistemic agent could safely infer that these practices are shared by nearly all other Hasidic Jews (at least those belonging to the same sect).[35] If I am right, then a social kind such as HASIDIC JEW provides us with an impressive metaphysical ground for inductive inference. These inferences seem to be much nearer in quantity and reliability to inferences afforded by biological and chemical kinds than they are to a kind like THE MORAL GOOD. Consequently, appeal to the existence of HPC social kinds fails to help the case for homeostatic consequentialism.

At this point, two complaints might be raised about my example. First, it might be complained that Hasidic Jews make up an unusually uniform social kind. According to this objection, the norms governing Hasidic life are more far reaching and pervasive than those governing other social groups. A more typical social kind would be much

---

[34] I am here offering these sample inferences after having done only a minimum of research. They are based on my own limited casual observations of Hasidic Jews (along with bits of testimony from others). No doubt, some of these observations need refinement or correction (for instance, I have not said—because I do not know—on which occasions tallis is worn). In any case, there can be little doubt that it would require only a modest amount of sociological research to extend both the number and the reliability of inductive inferences that can be made about Jacob in his capacity as a Hasidic Jew.

[35] It should be recognized that we should expect inferences to be reliable only when they concern properties that are homeostatically clustered. Suppose that my sample of Hasidic men was small, consisting of only five men. Suppose further that I observed that all five men have gray beards. I might be tempted to infer that Jacob's beard will be gray as well. It should be clear, however, that, even if this conclusion were to turn out to be true, this inference is not epistemically warranted. What this observation suggests is that, if our practice of making inductive inferences from a small sample is to be practical, then we had better have some skill at detecting which properties of an individual are essential to its kind. (For a defense of the claim that human beings really do possess such a skill, see Kornblith [1993: 83-107]).

nearer to THE MORAL GOOD in its (weak) grounding of induction. A second possible objection is that the uniformity among male Hasidum is artificial since it results from behaviors that individuals consciously undertake in order to remain members of the group. A genuine HPC social kind would not be unified by such an artificial mechanism.

Now, for those who subscribe to Boyd's HPC account of kinds, the "artificiality" of the homeostatic mechanism ought to be beside the point. After all, Boyd includes human artifacts like THE 1969 PLYMOUTH VALIANT as examples of HPC kinds (1999b: 68).[36] Nevertheless, I will offer one more illustration of a potential HPC social kind. The kind I will cite grounds a number of reliable inductive inferences but is not open to either of the above complaints. Its members are significantly less uniform than members of HASIDUM and do not (as far as I know) engage in their kind-typical behaviors for the express purpose of maintaining their membership within it. Consider, then, the social kind designated by the term 'hippy.' When we learn that Bill is a hippy we can reliably infer the following propositions: Bill owns at least one tie-dyed shirt and at least one pair of sandals; he listens to (or at least can appreciate) the music of The Grateful Dead and Phish; he has smoked marijuana and supports its legalization; in politics he opposes aggressive foreign policy and socially conservative domestic policies. To be sure, the reliability of these inferences will be much weaker than the inferences involving biological kinds; there are certainly many more hippies that do not enjoy the music of Phish than there are scorpions without claws. Still, I suspect these inferences are reliable

[36] It is worth adding here that Ron Mallon (2003) defends an HPC conception of certain social kinds where the homeostatic mechanisms include the members' own recognition of themselves as members of a kind. .

enough to meet the threshold for epistemic warrant. We see, then, that even a social kind like HIPPY grounds significant inductive inference where THE GOOD does not.[37]

The comparison of THE GOOD with plausible examples of HPC social kinds reveals once again that THE GOOD supplies us with a very weak metaphysical ground for reliable inductive inferences. Since one of the most notable features of HPC kinds (and natural kinds more generally) is supposed to be the role they play in grounding reliable inductive inferences, this observation supports the conclusion that THE GOOD is not itself an HPC kind.

Of course, I have cited only two examples of HPC social kinds. The homeostatic consequentialist may hold out in the hope that some social kind will be found that both (*a*) weakly grounds inductive inference and (*b*) is a convincing example of an HPC kind. (Perhaps it will be thought that one of the other social kinds I list at the start of this section can fit the bill.) I think such a hope is misplaced. Conditions *a* and *b* are in tension with one another. To the extent that a kind grounds inductive inference only very weakly, there will be good reason to doubt it is an HPC kind. If so, then we should not expect to find any convincing examples of an HPC social kind that grounds inductive inference as weakly as THE GOOD does.

[37] Some might object that the inferences about Bill are not inductive at all. It may be that 'hippy' has an analytic cluster definition where the properties I have attributed to Bill are just those that are analytically associated with 'hippy.' I am not inclined protest very loudly against this objection. But note that this should be of no comfort to the homeostatic consequentialist. After all, the same objection may be raised against any (supposedly) inductive inference involving THE GOOD (i.e., it might be objected that such inferences are not examples of *a posteriori* inductions at all, but are, instead, examples of the analytic *a priori*). In any case, unless a clear case of a HPC social kind that weakly grounds inductive inference can be found, the homeostatic consequentialist cannot appeal to a comparison with social kinds in order to answer the arguments of §5.5.2 and §5.5.3.

5.6. An Anticipated Rebuttal.

In a more recent presentation of his ethical views, Boyd suggests that the HPC that constitutes *goodness* is currently "fragmented" (2003b: 34-38). It might be thought that this claim provides the homeostatic consequentialist with a reply to the arguments of §5.5. In this section, I consider the prospects for such a reply.

As we saw in §5.3, Boyd suggests that several of the mechanisms that hold the human goods in a homeostatic relationship are socio-political. They include democratic institutions, social egalitarianism, certain customs etc. In different social environments, these mechanisms may be stronger, weaker, or even completely absent. Boyd's view is that as these mechanisms are made stronger—as a society becomes more democratic, more egalitarian, etc.—the human goods will become more strongly homeostatically unified (*ibid*.). As I understand it, this amounts to the claim that, as the relevant homeostatic mechanisms are strengthened, there will be an even greater increase in the likelihood than before that, when some of the human goods are instantiated, the other goods are instantiated as well.

Boyd submits that, at present, the sorts of mechanisms expected to produce homeostatic unity among the human goods are not nearly as strong as they could be:

> So far we have always operated morally within social structures which lacked the resources (technical or social or economic or political) to achieve the sort of (homeostatic) unity of [*goodness*][38] towards which our moral concerns aim, and which possessed lots of features "designed" as it were (often literally designed) to prevent the emergence of such resources (2003b: 36).

He goes on to suggest that, because of the poor present state of the relevant social institutions, the HPC that constitutes *goodness* is "not now very unified" (*ibid*.). This claim might be thought to supply homeostatic consequentialism with a reply to the

[38] Boyd uses 'the good' here.

arguments of §5.5. Here is how. If *goodness* is constituted by a weakly or partially unified HPC, then we should expect that the presence of some of the goods only slightly raises the likelihood that the others are present. In that case, however, we should not expect to find that reference to THE GOOD facilitates many reliable inductive inferences. These considerations show that it is possible that *goodness* is an HPC even though it fails to ground the sorts of reliable inductive inferences that are characteristic of HPC kinds. In light of this possibility, it might be argued that we are unjustified in concluding that *goodness* is not an HPC from the fact that it fails to ground significant inductive inferences.

It is difficult to assess this reply without a more detailed account of property homeostasis and of what sorts of characteristics a property cluster must have in order to serve as the real essence of a natural kind. It is not obvious that just any amount of homeostatic clustering among a group of properties is sufficient to make that group suited for the role of natural kind's real essence. Indeed, as we saw in §5.2, Boyd takes it to be characteristic of HPC kinds that the clustering of their defining properties is "causally important." However, to the extent that the purported clustering of the human goods fails to make a noticeable difference to the inductive inferences we are licensed to draw, it would seem that such clustering is not all that causally important. In addition, it should be recalled that Boyd himself takes it to be a defining feature of natural kinds (and so, HPC natural kinds) that reference to such kinds facilitates explanation and inductive inference. If he is right, then the fact that reference to THE GOOD fails to ground inductive inference in any interesting way should be thought to give us very strong grounds for

denying that it is itself an HPC kind even if it should so happen that the human goods are weakly unified.

Suppose, however, that we grant that a weakly unified cluster of properties is capable of playing the role of a natural kind essence. It still remains the case that the lack of interesting reliable inductions yielded by reference to THE GOOD leaves us with no assurance that the human goods, or any other collection of properties suitably related to the predicate 'morally good,' really are in fact weakly unified. This lack of assurance might not worry the homeostatic consequentialist. While he cannot confirm his empirical hypothesis (that the goods are weakly unified), he might suppose that the burden of proof is on his detractors to show that this empirical hypothesis is false. I think this stance would be a mistake. To see why, we need to revisit the dialectic.

In "How to Be a Moral Realist," Boyd raises the possibility that the human goods are homeostatically clustered and that this cluster can be identified with the property *moral goodness*. This hypothesis is meant to keep alive the possibility that some naturalistic version of moral realism is true despite a battery of anti-realist and non-naturalist objections. I take the arguments of §5.4 above to show that, even if the human goods that Boyd cites really do form an HPC, we should not identify this HPC with *moral goodness* itself. My own view is that the arguments of §5.4 are sufficient to refute homeostatic consequentialism outright. More cautiously, however, I would insist that those arguments at least deprive the theory of any presumption of innocence it might have enjoyed and give us at least some positive reason to think it false. At this point in the dialectic, then, the balance of reasons is against the hypothesis that *moral goodness* is an HPC.

Perhaps the best evidence that could be adduced in favor of homeostatic consequentialism would be an observation to the effect that reference to *goodness* or THE GOOD facilitates reliable inductive inference. Unfortunately, the arguments of §5.5 show that there is at present little reason to believe that reference to *goodness* or the good does in fact facilitate reliable inductive inference. This finding constitutes more than a mere failure to uncover exculpating evidence in favor of homeostatic consequentialism. In light of Boyd's claim that it is definitive of natural kinds (and thus, HPC kinds) that such kinds ground inductive inferences, the findings of §5.5 give us additional evidence for the denial of the proposition that *goodness* is an HPC. As I see it, then, the score is now (at least) 2 to 0 against homeostatic consequentialism.

It is at this point that the hypothesis that the HPC of human goods is weakly unified becomes relevant. This hypothesis promises to explain why it would be that *goodness*, though an HPC, fails to ground reliable inductive inference. If this hypothesis turns out to be consistent with the general theory of HPC kinds, and if we allow that our current evidence does not rule out this hypothesis, then the conclusion of §5.5 must be weakened. We could no longer take the lack of reliable inductive inferences afforded by reference to the good as positive evidence for the denial of the claim that *goodness* is an HPC. Instead, the observations of §5.5 should be taken to show merely an absence of evidence in favor of the affirmation of that claim. But this is too little too late. At best, the homeostatic consequentialist gets to turn the scoreboard back to 1 to 0; but he is still losing. In light of the arguments of §5.4, the balance of reasons still favors the denial of the proposition that *goodness* is constituted by an HPC.

5.7. Conclusion.

I have argued that the homeostatic property cluster account of *goodness* is false. First, it fails to account for the value of "causally isolated" human goods. Next, the relationship between THE GOOD and its defining properties is suspiciously unlike the relationship between paradigmatic HPC kinds and their defining properties. Finally, reference to THE GOOD does not support inductive inference nearly as well as would be expected if it were an HPC kind. For these reasons, we should conclude that *goodness* is not an HPC and that THE GOOD is not an HPC kind. As it stands, then, ethical naturalists remain without justification for adopting moral semantic externalism.

CHAPTER 6

THE EXPLANATORY IMPOTENCE OF MORAL FACTS

6.1. Introduction.

In Chapter 1 we saw that the main proponents of SEN endorse the method of reflective equilibrium as the proper way to conduct moral inquiry. This method directs an epistemic agent to make modifications to her moral beliefs about particular cases, general moral principles and theories, and non-moral background beliefs until these three elements exhibit maximal coherence. When these elements do exhibit maximal coherence, the agent's moral beliefs can be said to be in reflective equilibrium. In this state, her beliefs enjoy maximal epistemic justification. Call the moral theory that would survive the method of reflective equilibrium for an agent (or a group of agents), S, S's *best* moral theory.

Among those who endorse the method of reflective equilibrium for moral inquiry, there is disagreement about the metaphysical commitments of our best moral theory. On the one hand, there are those, such as the defenders of SEN, who embrace a realist construal of moral theory (Boyd 1988; Brink 1989; Daniels 1979; Sturgeon 2002). For the moral realist, our best moral theory should be thought of as stance-independently true. Moral anti-realists, by contrast, deny that our best moral theory should be thought of as stance-independently true. Some anti-realists deny that our best moral theory should be thought of as true at all. They argue, instead, that we should think of our best theory merely as a useful fiction (Mackie 1977: ch. 5; Joyce 2001; Nolan *et al*. 2005). Other moral anti-realists suggest that our best theory should be thought of, not as true, but as

"reasonable for us to accept" (Rawls 1980: 570). Still other moral anti-realists are willing to view our best moral theory as true, but maintain, contra realists, that that the theory's truth is stance-dependent. Anti-realists of this stripe hold that the truth of our best moral theory *consists* in the fact that it is the theory that we would accept, were our beliefs in reflective equilibrium (Rawls 1980: 519). This kind of metaethical view amounts to the adoption of a coherence theory of truth for moral claims.

What sorts of considerations might help us decide between the moral realist's construal of reflective equilibrium and the moral anti-realist's construal? On the one hand, it may seem as though phenomenological considerations favor moral realism. When we engage in moral deliberation, it typically seems to us as though there is a correct answer, independent of what we happen to believe, that we are trying to arrive at. Furthermore, it seems to us as though our best efforts might fail to yield the correct answer. More specifically, it seems to us logically possible that we achieve reflective equilibrium among our beliefs and yet the moral theory that we accept is false while some other moral theory that we reject is true (Brink 1989: 31-36).

Anti-realists have countered that the phenomenological evidence for a realist construal of moral inquiry is defeated by the "explanatory impotence" of would-be moral facts, where a fact is explanatorily impotent just in case it is not needed in the best *a posteriori* explanations of our observations (Harman 1977: 3-23). Indeed, the putative explanatory impotence of moral facts is taken not only to defeat the presumptive evidence in favor of moral realism, but also—in light of Ockham's razor-type considerations of ontological parsimony—to constitute positive evidence against a realistic construal of moral inquiry.

Against the argument from explanatory impotence, some moral realists have objected that it begs the question against moral realism to suppose that moral facts must play a role in our *a posteriori* explanations. Nagel writes, "The claim that certain [e.g. moral] reasons exist is a normative claim, not a claim about the best causal explanation of anything" (Nagel 1986: 144; cf. Quinn 1986; Shafer-Landau 2003: 98-114; 2006). Whatever the merits of this line of reply, it is not available to proponents of synthetic ethical naturalism. Recall from Chapter 1 that metaphysical naturalists accept the following explanatory criterion of ontological commitment:

> EC: Posit the existence of an entity (or a kind of entity) if and only if reference to that (kind of) entity is needed in our best available *a posteriori* explanations of observable phenomena.

If stance-independent moral facts are to be a welcomed part of the metaphysical naturalist's ontology, then reference to such facts had better be needed in our best available *a posteriori* explanations. In light of their commitment to EC, it is not surprising to find that synthetic ethical naturalists respond to the argument from explanatory impotence by insisting that, contrary to what anti-realists have claimed, moral facts really do figure in our best *a posteriori* explanations (Boyd 1988; Brink 1989: 182-197; Sturgeon 1985a).[1]

In this chapter, I defend the moral anti-realist's argument from explanatory impotence against the naturalists' rebuttal. In §6.2 below, I present Gilbert Harman's argument from explanatory impotence along with Nicolas Sturgeon's *tu quoque* reply to it on behalf of SEN. In §6.3, I outline a revised version of the argument from explanatory impotence that I believe to be invulnerable to Sturgeon's objections. In §6.4, I consider

[1] Although Brink unequivocally affirms that moral facts explain our observations, it should be noted that he expresses some ambivalence about whether they really must do so in order to be a legitimate part of the naturalist's ontology (Brink 1989: 182f).

several replies to the revised argument from explanatory impotence. I pay special attention to another *tu quoque*-style argument to the effect that the reasoning behind the explanatory impotence argument against moral realism, if cogent, would commit us to the rejection of scientific realism. In §6.5, I argue that scientific realism is not vulnerable in this respect. (In the next chapter, I go on to show that the sort of argument discussed in §6.5 here, which protects scientific realism from explanatory impotence worries, cannot be extended to defend moral realism in a similar fashion.)

6.2. The Harman-Sturgeon Exchange.

6.2.1. Harman's opening salvo.

The contemporary *locus classicus* for the argument from explanatory impotence is Gilbert Harman's *The Nature of Morality* (1977). There, he argues that putative moral facts are not needed in order to explain why we have the moral beliefs that we do. To illustrate, he has his readers imagine a case in which they observe "a group of young hoodlums pour gasoline on a cat and ignite it." Upon making this observation, we are to imagine that we (the observers of the act) form an immediate judgment that the hoodlums' act is morally wrong. Harman asks whether we need to suppose that the act of igniting the cat really does have the property of *being morally wrong* in order to explain the fact that we made this moral judgment. He answers in the negative:

> …[A]n assumption about moral facts would seem to be totally irrelevant to the explanation of your making the judgment that you make. It would seem that all we need to assume is that you have certain more or less well articulated moral principles that are reflected in the judgments you make, based on your moral sensibility. It seems to be completely irrelevant to our explanation whether your intuitive immediate judgment is true or false (Harman 1977: 7).

To explain our judging the hoodlums' act to be morally wrong, we need to cite only: (a) our non-moral beliefs about the relevant act-token (e.g., the fact that we believe these young people to be lighting the cat on fire, that they have done this merely for their own amusement, and that it is causing the cat to experience a great amount of pain) and (b) the fact that we (perhaps tacitly) accept a moral principle according to which acts of causing suffering for mere amusement are morally wrong. I suspect, in addition, that in order to adequately explain our having the non-moral beliefs that we do about the action, we will also need to cite (c) the non-moral facts about the act-token itself (e.g. the fact that there really are young people lighting a cat on fire). The conjunction of these three items forms what is at least a plausible and satisfying explanation of why we judge the act of lighting the cat on fire to be morally wrong.

Some might object that an even better explanation would describe these same items at a more fundamental level of reality, perhaps at the level of physical particles; but this objection is quite compatible with the general outlines of Harman's argument. What is most important for the success of the explanatory impotence argument is that the explanation of our moral judgment would not be improved by expanding it to include a claim to the effect that the act of lighting the cat on fire really is morally wrong. Since an explanation that omits reference to moral facts or properties has the theoretical virtue of being more ontologically parsimonious than any explanation that does make such a reference, Harman's non-moral explanation of our judgment would seem to be *ceteris paribus* better than any competing moral explanation (i.e., an explanation that makes ineliminable use of moral vocabulary). Unless there is some reason to think that non-moral explanations of our moral beliefs are inferior to moral explanations in some other

respect, a non-moral explanation of the sort described above would seem to be the best available to us. Harman's view is that moral explanations are not superior to non-moral explanations in any significant respect.

It appears, then, that moral facts are not needed in the best explanations of our making this or that particular moral judgment. If we are metaphysical naturalists, and, thus, we accept a principle like EC, then it would seem to follow that we should not believe in the existence of stance-independent moral facts. The upshot is that metaphysical naturalists should deny moral realism.

6.2.2. Sturgeon's reply to Harman.

Over the course of several papers, Nicholas Sturgeon mounts a defense of naturalistic moral realism against Harman's explanatory impotence argument.[2] There are two major components of Sturgeon's defense. First, Sturgeon presents several cases in which a moral explanation of a non-moral fact appears to be both plausible and not obviously inferior to any available rival explanation.[3] He suggests, for instance, that the fact that Hitler was morally depraved forms part of a good explanation for why we believe that he was depraved. Not only does Sturgeon claim that moral facts sometimes figure in (what are potentially) the best explanations of our moral beliefs, he argues that they also figure in the (potentially) best explanations of non-doxastic events or states of affairs, such as the fact that a given person has performed a particular action. So, for example, Sturgeon suggests that the fact that Hitler was depraved forms part of a reasonable (and potentially

[2] Sturgeon's most direct replies to appear in his (1985a) and (1986a). Other Sturgeon papers that are relevant to the moral explanations debate are his (1992), (1998), and (2006).
[3] In his (1985a: 56), Sturgeon's stated goals are modest. He does not claim show that moral facts are *in fact* needed in our best a posteriori explanations, but only that we do not now know that they are not (or will not be) needed.

best) explanation of why Hitler performed the sorts of actions for which he is commonly reviled, such as ordering the extermination of European Jews.

I want to delay an in-depth examination of this component of Sturgeon's defense until Chapter 7 (see especially §7.6). Here, I will mention one reason that the latter Hitler example is not very compelling as a potentially best explanation of non-moral facts. There are a number of natural properties that could realize the property of being morally depraved: e.g., *being homicidal*, *being dishonest*, *being sadistic*, *being a pedophile*, *lacking empathy*, etc. The proposition that Hitler is morally depraved does not, by itself, inform us as to which of these putative *depravity*-making properties he has. In light of this, it is doubtful that Hitler's being depraved helps in explaining his actions. If Hitler's depravity had been realized by, say, his being a pedophile, rather than by his being homicidal, or being sadistic, or lacking empathy, it is not likely that he still would have ordered the extermination of European Jews (even assuming that he would be in a position of power to do so).[4] In light of this, it would seem that a satisfactory explanation of Hitler's actions cannot simply cite the fact that he was morally depraved; rather, it must ascribe to him a particular depravity-making natural property. However, since true ascriptions of these latter sorts of properties (e.g., *being homicidal*, *being sadistic*, etc.) are compatible with the falsity of all moral theories, it would seem that the best explanation of Hitler's actions in terms of his character traits need not make reference to

[4] One might complain that, the social-political situation of Germany being what it was in the 1930's, even a non-homicidal, non-sadistic Hitler with normal powers of empathy would have ordered the final solution anyway. But if that is true, then the natural thing to conclude is that Hitler's depravity played no role in causing his notorious actions. But this conclusion, of course, is exactly the contradictory of what the naturalist is trying to establish.

any distinctly moral properties or facts; or, to put a linguistic spin on what is essentially the same point: such explanations can be expressed without using moral vocabulary.[5]

6.2.3. Sturgeon's *tu quoque* reply.

To my mind, Sturgeon's second response to Harman's explanatory impotence argument is more promising. Sturgeon observes that an argument parallel to Harman's can be constructed to support an anti-realist construal of scientific theories. In his original presentation, Harman contrasts the explanation of an observer's moral beliefs with a case involving another observer forming a scientific belief. In this contrasting case, a physicist observes a vapor trail in a cloud chamber. Upon making this observation, the physicist forms the belief that a proton has just passed through the chamber. Harman contends that, in contrast to the moral case, the fact that a proton passed through the cloud chamber really does constitute part of the best available explanation of the physicist's belief. This contrast between the moral and scientific case is meant to highlight the trouble with moral facts: whereas we need to assume the existence of theoretical facts (e.g., the fact that a proton has passed through the chamber) in order to explain the physicist's judgment, we do not need to assume the existence of moral facts in order to explain our making a moral judgment about the torturing of the cat.

[5] The reason it may be better to speak of explanations that do not use moral vocabulary, as opposed to explanations that do not make reference to moral facts, is that the latter way of speaking leaves the anti-realist vulnerable to the charge of begging the question against the ethical naturalist. Intuitively, an explanation making ineliminable reference to a natural property like *maximizing the balance of pleasure over pain* should not be counted as a moral explanation (*ceteris paribus*). But ethical naturalists like Brink and Sturgeon have claimed that it may well turn out that this natural property is identical with *moral rightness*. If such an identity claim were true, it would follow via Leibniz' Law that the explanation in question needs to make reference to a moral property after all (viz., the moral property *maximizing the balance of pleasure over pain*). In that case, we could not say, without begging the question against the naturalist, that the relevant explanation makes no reference to moral properties or facts. To avoid the charge of begging the question, then, it may be more appropriate to speak of moral facts as being explanatorily impotent in the sense that the best explanations of our observations can be stated without the use of moral vocabulary.

Against this, Sturgeon notes that an analogue of Harman's explanatory impotence argument would show that the proton is not needed to explain the physicist's belief after all:

> Given her training and the theory that she has internalized, the physicist would have thought, "There's a proton," at the sight of a vapor trail, whether the trail had been produced by a proton or not: so it looks like assumptions about the proton are *not* needed, after all, to explain her observational judgment (2006a: 245f; cf. Sturgeon 1985a: 68-71; Brink 1989: 184-186).

With respect to Harman's physicist example, then, it would seem that we can explain the fact that the physicist believes that a proton has passed through the chamber simply by citing (a) the physicist's non-theoretical, observational beliefs (e.g. her belief that there is a vapor trail in the cloud chamber), (b) the fact that she accepts a physical theory according to which vapor trails indicate protons under these conditions, and, perhaps, (c) the fact that there actually was a vapor trail in the cloud chamber (this explains her observational belief).[6] If Harman's argument from explanatory impotence is sufficient to motivate the rejection of moral realism, then the availability of this non-theoretical explanation of the physicist's belief ought to be sufficient to motivate the rejection of realism about protons. Since, however, all parties to this debate accept realism about theoretical entities like protons, the availability of this parallel argument shows that Harman's argument against moral realism must be defective somehow. It would seem,

[6] Some have raised the question of whether the third item, c, is really necessary in order to explain the physicist's belief. After all, given the theory that she accepts, the physicist would have judged that a proton passed through the chamber even if she had only experienced a visual experience of a vapor trail, but no real vapor trail was present. Perhaps this is right. Whatever the case may be, it isn't of great importance for the point being made. Nevertheless, it seems to me that, if the physicist did have a visual experience of a vapor trail, this is something that cries out for explanation. One possible explanation is that she is a brain in a vat and that the computer generating her visual images has run a program that caused her to experience an appearance of a vapor trail. Another possible hypothesis is that, due to spending long nights overworking at the lab, she has simply suffered a hallucination of a vapor trail. Above, I recommend the hypothesis that there really is a vapor trail because this seems to me the most plausible explanation for the majority of actual-world cases in which a physicist has a visual experience of a vapor trail. But again, whether or not a real vapor trail turns out to be part of the best explanation for her belief, the point being made stands.

then, that the explanatory impotence argument must be rejected (at least by any philosopher who accepts scientific realism, such as me).

With respect to this particular exchange, I am inclined to think that Sturgeon comes out on top. But, as we will see, there is still good reason to doubt that that moral realism is safe from the charge that putative moral facts are explanatorily impotent.

6.2.4. Lessons of the Harman and Sturgeon exchange.

In his initial presentation of the argument from explanatory impotence, Harman selects an individual's arriving at a particular moral judgment as the explanandum for which moral facts are potential explanans. It is this choice that makes his argument vulnerable to Sturgeon's *tu quoque* reply. Harman is right to think that the best proximate explanation of why some moral appraiser arrives at a particular moral judgment about an action will need to make reference only to facts about the moral theory she accepts along with other, non-moral facts about the case. As we saw, however, similar things are true *mutatis mutandis* of the proximate explanations of why a physicist arrives at the judgment that a proton has passed through a cloud chamber: the best proximate explanation of her judgment will need to cite only facts about what scientific theory she accepts along with other, non-theoretical facts about the case (e.g., that she observed a vapor trail, etc.).

The lesson for the moral anti-realist is that a successful explanatory challenge to moral realism will have to investigate more *ultimate* explanations of our moral judgments. The question that needs to be asked is not whether moral facts figure ineliminably in the best explanation of our having made this or that *particular moral judgment*, but rather, whether moral facts figure in the best available explanations of our

having the *moral sensibility* that we have.[7] In this context, we can think of a person's moral sensibility as a (usually inchoate and tacitly held) moral theory or standard that she accepts.[8] If moral realism is true, and if we possess approximate moral knowledge, then we should expect that stance-independent moral facts play an important role in the best explanation of our having the moral sensibility that we currently have. If moral facts are not needed in such explanations, then we should either reject moral realism or else accept the skeptical view that we have no moral knowledge.

Notice that naturalistic moral realists themselves acknowledge this challenge, or something like it. Boyd, Brink, and Sturgeon all acknowledge that, if the method of reflective equilibrium is to be thought of as a reliable method that guides us towards true moral belief, then it had better be the case that our initial stock of pre-theoretical moral beliefs are at least "approximately true":

> If a dialectical process [of moral reasoning] that takes common or considered moral judgments as a significant part of the input is to have moral knowledge as output, then there ought to be reason to think that the judgments of commonsense

---

[7] Sturgeon appears to recognize that this is the location of an important challenge to moral realism when he writes, "…I think the main problem [for moral realism] arises only when we take a moral global view of the history of moral and scientific thought" (1986a: 70f; cf. 1992: 101; 2006: 254f; cf. Quinn 1986). In some places, Sturgeon suggests that this challenge is related to worries about the existence of deep moral disagreement and the difficulty of settling such disagreements (see, for instance, his 1985a: 49). I think there is something to this. But it is important to see that the present challenge could be posed even if it so happened that everyone in the world shared approximately the same moral beliefs. Even if there were such agreement, we could still ask whether stance-independent moral facts constitute an ineliminable part of the explanation for the convergence in moral sensibility (cf. Williams 1985: Ch 8). Suppose, for example, convergence in moral sensibility has been achieved, not through argument and presentation of evidence, but rather, through the repression and extermination of those with dissenting moral sensibilities. If this were our own situation, I suspect that most would agree the fact that the people agree in their moral sensibilities constitutes very little evidence in favor of moral realism. (Certainly, we would not think it good evidence for the truth of some religion, R, if the entire population of the world came to accept R unanimously as a result of coercion and extermination of dissenters.)

[8] It might be helpful to follow Simon Blackburn and think of a person's moral sensibility roughly as a function that takes non-moral observations or non-moralized descriptions of actions, states of affairs, (etc.) as input and yields moral judgments as output (Blackburn 1998: 5). Note that, just as a speaker of a language easily obeys grammatical rules that she may be unable to articulate, a moral appraiser need not be in a position to articulate the moral theory or standard that underwrites her moral sensibility (assuming, contra particularists, that there even is such a standard). In short, an appraiser may (indeed, will) often lack easy introspective access to the contents of her moral sensibility.

> morality are sufficiently close to the truth. Dialectical inquiry can identify and correct various sorts of errors, even very significant and far-reaching errors, but it appears unable to identify or correct systematic error, because the grounds and direction for correction must emerge from reflection on the beliefs with which cognizers start. Do we have any reason to think that the considered moral convictions of commonsense morality are generally reliable or at least not systematically seriously mistaken? (Brink 1999: 207; cf. Boyd 1988: 201, 207; Brink 1989: 299; Sturgeon 1986a: 67; 2006a: 254f.)

If moral facts themselves do not figure in the best explanations of our (past or present) moral sensibility then it is hard to see how we can answer Brink's question affirmatively. In that case, we should conclude either that moral skepticism is true (i.e., that we have no moral knowledge), or else that some form of moral anti-realism is true. Of course, an argument from explanatory impotence refocused in this way will represent an improvement over Harman's original version only if it is not vulnerable—or, at any rate, significantly less vulnerable—to another *tu quoque* reply.[9] After presenting a refocused version of the explanatory impotence argument against moral realism in §6.3 below, I will go on to argue in §6.4 and §6.5 that this argument is indeed much less vulnerable to a *tu quoque* reply. Before turning to these matters, however, I need to address a worry that my argument neglects a possible avenue open to naturalist moral realists.

6.2.5. Moral explanations of non-doxastic phenomena.

Above, I suggested that the central question for the moral explanations debate is whether moral facts are needed in the best explanations of our having the moral sensibility that we have. It might be thought that this neglects another possibility: perhaps moral facts are needed to best explain something other than our moral sensibility or moral beliefs. If

[9] The sort of *tu quoque* reply that I have in mind would involve presenting a parallel argument to the effect that theoretical facts play no needed role in the best explanations of why we accept the scientific theories that we currently accept.

so—if moral facts are needed in the best explanations of some non-doxastic phenomena—then by EC, metaphysical naturalists will be permitted (indeed, required) to accept the existence of moral facts *even if they are not needed to explain our moral sensibility*. Consequently, even if a good case can be made that moral facts do not explain our moral sensibility, it can be argued that it would be too hasty to conclude on the basis of EC that we should not posit their existence.

This line of defense can promise only cold comfort to moral realists, and perhaps not even that. In the first place, if moral facts do not figure in the best explanations of our moral sensibility and moral beliefs, then, given some fairly orthodox epistemological assumptions (e.g., an anti-luck condition on epistemic warrant), it follows that we have no moral knowledge. This skeptical conclusion follows even if it should happen that moral facts really do figure in the best explanations of some non-doxastic phenomena somewhere in the universe.

Now, moral skepticism (the view that we have no moral knowledge) is not logically incompatible with moral realism as I have formulated it in §1.2.1: it is coherent to suppose that there exist stance-independent moral facts and that human beings have no knowledge of these facts. Still, a commitment to moral skepticism would surely be a disappointment to the general metaethical outlook of moral realists; realism is typically presented as part of a non-skeptical view of morality. Indeed, in their own formulations of moral realism, both Sturgeon and Boyd include an anti-skeptical epistemological condition. Sturgeon, for example, writes that a "core thesis" of moral realism that "…our ordinary methods of arriving at moral judgments provide us with at least some approximate knowledge of moral truths" (Sturgeon, 1986b: 117; cf. Boyd 1988: 182).

A commitment to moral skepticism threatens more than mere disappointment to moral realists, however. If we have no moral knowledge, then it is doubtful that we currently know that there are phenomena that moral facts are needed to explain—for if we did know this, then presumably we would have *some* moral knowledge. But if we are aware that we do not know that there are phenomena that moral facts are needed to explain, then, if we accept EC, we are not within our rights to posit the existence of moral facts. Thus, even if our evidence fails to entail that moral realism is false, the fact that we lack moral knowledge does seem to entail (at least, when conjoined with EC) that we ought not now accept moral realism.

In addition to these worries, there is also the problem of finding plausible non-doxastic explanatory work for moral facts to do. If moral facts do not explain our moral beliefs and moral sensibility, what other phenomena might there be left for them to explain? One plausible suggestion is that moral facts explain human actions. Indeed, we do sometimes say of a philanthropist, for example, that she donates to charity because it is the morally right thing to do. However, in paradigmatic cases of this sort, we expect that the putative moral fact will move the agent to act only if she believes that fact. More concretely, we expect that the fact that donating to charity is morally obligatory will move the philanthropist to act *only insofar as the philanthropist believes that donating to charity is morally obligatory*.[10] Thus, even if moral facts explain human actions, such explanations will be plausible only if those facts also explain our moral beliefs and sensibility. So we are still in search of non-doxastic phenomena that moral facts might

[10] It is not out of the question that moral facts could influence an agent to act by some kind of subconscious process, bypassing her beliefs and moral attitudes. But surely this would not be an attractive hypothesis if we had no evidence that moral facts ever cause an agent to act because she consciously apprehends that such a fact. It would be very strange indeed if moral facts exert a causal influence on our actions but *only* through subconscious processes.

plausibly explain. Unfortunately, outside of human beliefs and human actions, there just aren't any phenomena remaining for which facts about what is morally right or wrong, good or bad, etc., might plausibly be needed to explain: moral facts of these sorts are certainly not needed to explain planetary motion, plate tectonics, weather patterns, the behavior of atomic particles, etc. If this is right, then assuming EC, the case for naturalistic moral realism really does hang on the question of whether putative moral facts are needed in the best explanations of our moral beliefs and moral sensibility.

6.3. Anti-Realist Explanations of Moral Theory.

6.3.1. Non-moral explanations.

I have suggested that a better version of the moral anti-realists' argument from explanatory impotence would focus on explanations of our having the sort of moral sensibility that we have, rather than on proximate explanations of our making this or that particular moral judgment. (It will be useful to speak of an entire community as having a moral sensibility or accepting a moral theory. Presumably, the matter of which moral theory a community accepts is some kind of function of facts about which moral theory [or moral theories] its individual members accept.)[11] The anti-realist's task is to advance a non-moral explanation of our accepting the moral theory that we accept. In this context, an explanation counts as non-moral if it does not make reference to stance-

[11] This idea of a community having a moral sensibility cries out for more elaboration. A natural question to ask, for instance, is what percent of a community's population must accept theory T in order for T to count as giving the content of that community's moral sensibility? What should we say when a community that accepts T is a sub-community of a larger community, the majority of which accepts a different, incompatible moral theory T′? What should we say if the moral experts in a community accept T, while laypersons accept T′? I must leave these questions unanswered, since I simply do not know how best to answer them. My hope is that the notion of a community's accepting a moral theory is intuitive enough to utilize anyway. In an effort to avoid some of these difficulties, I recommend that we confine our attention to the moral sensibility that characterizes secular members of post-industrial, Western countries.

independent moral facts—or, at least, if it makes no use of moral vocabulary. If a non-moral explanation of this kind is to play a role in an argument against moral realism, then it must be superior to all realism-friendly moral explanations of moral theory that are on offer. That is, it must be superior to all competing explanations that make ineliminable reference to moral facts (or that make ineliminable use of moral vocabulary).

There are a number of anti-realist, non-moral explanations of moral theory that have been proposed. Some have long pedigrees. In the first place, there is an old line according to which the moral theory that is current in any given society is a mere reflection of the interests or preferences of the powerful in that society. This sort of view makes an early appearance in the mouth of Thrasymachus in Plato's *Republic*:

> …[T]he just is nothing else than the advantage of the stronger. […] …[E]ach government makes laws to its own advantage: democracy makes democratic laws, a despotism makes despotic laws, and so with the others, and when they have made these laws they declare this to be just for their subjects, that is, their own advantage, and they punish him who transgresses the laws as lawless and unjust (Plato 1974: 13 [338c-e]).

Although Thrasymachus identifies the property of *being a just law* with the property of *being a law that is to the advantage of the strong*, the thesis that is of interest here is his claim that the content of any polity's principles of justice is determined by whatever considerations the powerful in that polity deem to be to their own advantage. (Note that accepting this latter claim does not commit one to Thrasymacus's identity claim.) The same kind of view is also advanced by Marx and Engels' in *The German Ideology*:

> The ideas of the ruling class are in every epoch the ruling ideas: i.e., the class which is the ruling *material* force of society, is at the same time its ruling *intellectual* force. The class which has the means of material production at its disposal, has control at the same time over the means of mental production, so that thereby, generally speaking, the ideas of those who lack the means of mental production are subject to it. […] For each new class which puts itself in the place of the one ruling before it, it is compelled, merely in order to carry through its

> aim, to represent its interest as the common interest of all the members of society, that is, expressed in ideal form: it has to give its ideas the form of universality, and represent them as the only rational, universally valid ones (1933/1978: 172, 174, emphasis in the original).

The sort of picture suggested by Marx and Engels is one in which the moral standard propagated by the thinkers of a given society is the standard that best (or, at least, largely) serves the interests of their social class. Since only those thinkers that belong to the ruling class are in a position to propagate these ideas on a large scale, the moral standard that is accepted by a given society will always be reflective of the interests of the ruling class. It is not clear whether Marx and Engels suppose that the adoption of a standard by thinkers is a cynical plot. A more plausible story would have it that their selection of these principles is sincere, a mere result of their projecting their personal interests as they understand them onto the rest of their compatriots or onto their country as a unified whole.

If, as Thrasymachus and Marx and Engels suggest, the content of the moral principles or the tacit moral theory that we accept really is fixed by whatever happens to be in the interests of the powerful in our society, then it would seem to be unnecessary for us to invoke stance-independent moral facts in the explanation of why we accept those principles. This is so, at least, provided that there is no problem with the supposition that the thinkers of a society project their interests into putatively universally applicable moral rules.

A similar (although perhaps less crude) kind of non non-moral explanation of moral sensibility can be found in the writings of Freud and Nietzsche (Leiter 2001: 83-85). On the Freudian picture, roughly, each person's moral conscience (i.e., moral sensibility) arises as a result of her internalizing both the rules and expectations that her

parents (or some other external authority) imposes upon her, and the punishments expected as a result of breaking these rules (cf. Freud 1931/1961: 83-96; 1933/1965: 71-100; cf. Nietzsche 1887/1967: 84f). Again, genuine moral facts are not needed in this picture. Thus, if the Freudian explanation of moral sensibility is the best available, then by EC we ought to deny the existence of stance-independent moral facts.

6.3.2. A Darwinian account of moral sensibility.

Whatever the merits of the aforementioned anti-realist explanations of moral sensibilities, it seems to me that the most promising kind of non-moral explanation is of a different sort. In recent years, a number of philosophers have looked to Darwinian natural selection in order to explain the content of our moral sensibility. It should be noted at the outset that this sort of explanation is quite compatible with the claim that the sorts of considerations mentioned above exert an important influence the content of our moral sensibility. One advantage of the Darwinian account that is worth noting, however, is that it offers a plausible explanation not only of the content of our moral sensibility, but also of the origin of our concept of *moral obligation*. In the remainder of this section, I offer a brief sketch of a kind of Darwinian account of our moral sensibility.

To start, it must be observed that individual organisms can enhance their reproductive fitness[12] by cooperating with other individuals. This is readily evident with respect to predatory species that hunt prey that is too strong, too quick, or too endurant for a single predator to capture. Under certain circumstances, predators that cooperate

[12] There is some disagreement among philosophers of biology concerning how the term 'fitness' should be understood. For our purposes, Alan Gibbard's rough definition should suffice: "An organism's *fitness* is its expected degree of reproductive success, given its characteristics and its environment", where "[a]n organism's reproductive success is roughly the number of descendants it has in the distant future" (Gibbard 1990: 62).

will tend to be significantly more successful on hunts than their more solitary conspecifics. Greater success in hunting leads to greater access to nutrition and to a greater likelihood of survival. In turn, this gives the predator more time to breed, and a superior ability to provide nutrition for offspring; and these lead to a greater likelihood of leaving behind more surviving progeny than would otherwise be possible. Assuming that tendencies to behave cooperatively are heritable, we can expect that the genes responsible for cooperative behavior will be even more widely represented in each subsequent generation of this predatory species.

Two of the mechanisms most commonly cited in evolutionary explanations of cooperation are *kin selection* and *reciprocal altruism*. In cases of kin selection, cooperative behaviors (and even "self-sacrificing"[13] behaviors) are directed at closely related family members. To the extent that such behaviors result in a number of closely related family members enjoying greater reproductive fitness, the genes of the sacrificing individual will be passed on in greater numbers than would be the case if the individual had refrained from cooperating (Hamilton 1964a; 1964b; Ruse 1986/1998: 220). Where a tendency to cooperate arises between non-kin, it is standard to explain this by appeal to the mechanism of reciprocal altruism. In cases of reciprocal altruism, roughly, individuals help others at a cost to themselves, but with an "expectation"[14] that beneficiaries will reciprocate when the tables are turned (Trivers 1971). The reproductive benefits of cooperation accrue to those altruists who limit their helping only

[13] 'self-sacrificing' is in scare quotes here because, from a biological perspective anyway, it is not entirely clear that behavior of this sort constitutes a genuine sacrifice. A parent who dies protecting her young from predation, assuming her efforts succeed and her young go on to reproduce themselves, has arguably sacrificed nothing from the point of view of reproductive fitness (assuming she could not have lived long enough to replace the young that would have perished but for her efforts).

[14] In lower animals, we need not think of the individual's "expectation" of reciprocity as a propositional attitude had by that individual. What is important is that the individual has a behavioral disposition to discontinue behaving altruistically when the altruism is not reciprocated.

to reciprocating individuals. Because cooperating with cheaters (individuals who do not reciprocate) imposes a cost in fitness to altruistic individuals, natural selection favors those altruists with an ability to detect and exclude (and even punish) such cheaters.

Note that, according to the account being sketched here, the proximate mechanisms eliciting altruistic behavior in non-human organisms are affective or conative states. At least, this is so in those organisms sophisticated enough to count as capable of having such states. What natural selection gives to these organisms, then, is something like a *desire* to care for one's kin or to cooperate with those who have helped in the past.

So far, I have discussed the evolution of cooperative behavior. I have not said anything about the emergence of a genuine moral sensibility. How might this evolutionary account be extended to explain the fact that we make moral judgments and accept moral principles? Several philosophers advancing evolutionary explanations of moral thought have suggested that distinctly moral sensibilities evolved in large part as a solution for combating the individuals' temptation to defect from cooperation (Kitcher 1998: 302ff; Joyce 2006: ch. 4; Ruse 1986/1998: 221ff). The background for this suggestion is this: Even individuals who are generally cooperative have egoistic desires that compete with their altruistic desires and concerns. Presumably, this is because an individual with a mixture of altruistic and egoistic desires and dispositions enjoys an even greater reproductive fitness than individuals who behave in a *purely* altruistic way in all circumstances (Kitcher 1998: 299-302). However, the existence of egoistic desires in an individual will sometimes tempt her away from reciprocating, even in those circumstances in which it would be to her disadvantage to defect and act selfishly.

Indeed, an individual's failure to reciprocate could be quite costly to her as it may trigger nearby altruistic conspecifics to exclude her from future cooperative ventures and, in some cases, lead them to form an alliance against her.[15] Moreover, in groups in which defections from cooperation are common, the cooperative scheme will be less stable than that of groups whose members more reliably toe the party line. This instability both limits the size of workable cooperating groups and also results in a need for members to invest a large amount of time and effort into peacemaking (Kitcher 1998: 302f). Consequently, members of these groups lose out on advantages in reproductive fitness that could be theirs, but for a more effective cooperative scheme.

According to the evolutionary account of moral judgment, in order to combat failures to cooperate in the appropriate circumstances, our hominid ancestors evolved a moral sensibility. On this picture, an individual with a moral sensibility sees certain situations as *demanding* a certain response. To judge that one morally ought to cooperate is not simply to feel an inclination or desire to cooperate (as might be the case with non-moralizing altruistic animals); it is to see cooperation as something that is *required* of oneself irrespective of what one may happen to desire (Joyce 2006: ch. 2). In addition, it also involves seeing transgressors of moral rules as *deserving* condemnation or punishment (even if that transgressor is oneself). Important for the account being sketched here is the assumption that there is a strong link between the moral judgments that an individual makes and her motivation to perform (or refrain from) actions that are the subject of such judgments (Blackburn 1988: 363; Gibbard 1990: 76-80; Joyce 2006:

[15] Some of the philosophers who propose this style of evolutionary story point to studies of chimpanzee behavior for clues about the behavior of our primate ancestors. To the extent that present day chimp behavior really can give us some insight into the behavior of our ancestors, Frans de Waal's research provides us with evidence that our pre-moralizing ancestors formed cooperative alliances (see de Waal 1982).

108-118; Street 2006: 157n13).[16] What is being claimed, then, is that by seeing certain kinds of behavior as morally obligatory, individuals with moral sensibilities turn out to be more effective and reliable cooperators than non-moralizers; and, as a result, they enjoy greater reproductive fitness.

Something needs to be said now about what this account implies about *the content* of our moral sensibilities. What does this account imply about the matter of which tacit moral theory or principles we accept? No one who advances this evolutionary picture claims that the content of our moral sensibilities is wholly determined by natural selection. Most proponents do, however, claim that natural selection has influenced that content to a significant extent (Joyce 2006: 140; Ruse 1986/1998: 235-247; Street 2006: 113-121).[17] To begin with, note that, by this picture, altruistic and cooperative behavior develops before moral sensibilities emerge. As we saw, moral sensibility and moral judgment evolve as a way of making individuals more effective cooperators. In light of this developmental order, we should expect that the contents of the moral sensibilities of the earliest moralizers reflected the cooperative tendencies and preferences that were already prevalent among non-moralizing ancestors. For example, whereas our non-moral ancestors felt a strong desire or urge to feed and defend their young, our moralizing ancestors would, in addition, judge themselves to be under a moral obligation to do these things; they would see such actions as demanded by the situation; and they would see

[16] This need not—indeed, *should* not—be read as an assumption of the truth of moral judgment internalism. According to (one version of) moral judgment internalism, it is a necessary truth that any agent who judges herself to have a moral obligation to ϕ is to some extent motivated to ϕ. I am not assuming anything so strong here. All that is needed for the evolutionary story is the much weaker claim that *normal* moralizing agents, under *normal* conditions, are to a fairly strong extent motivated to perform the acts they judge themselves morally obligated to perform. Even naturalists who famously deny moral judgment internalism have seen fit to grant something like this weaker assumption (e.g., Boyd 1988: 215f; Brink 1989: 49).

[17] One outlier is Kitcher, who writes, "I have made no explicit claims about the emergence of morality from proto-morality, but it seems to me overwhelmingly plausible that this history has been guided mainly, if not exclusively, by forces of cultural, rather than natural selection" (1998: 305).

their own failure to comply with this demand as deserving condemnation. To the extent that non-moralized cooperative dispositions and behaviors were the result of natural selection, then, it is plausible to suppose that the content of the moral sensibilities of our earliest moralizing ancestors were greatly influenced by natural selection as well. Indeed, it should be observed that the basic contours of our own moral sensibility seem to favor kinds of behavior that would likely have enhanced the reproductive fitness of early humans, if not ourselves. To extend the previous example, we tend to judge that our moral obligations to aid our own children are much more stringent than our obligations to aid the children of strangers. It is not hard to see how agents who incorporate this judgment into their sensibility are likely to have greater reproductive success than those who do not.[18]

It is worth reiterating that this evolutionary account of the origin of our moral sensibility does not preclude the claim that other factors, such as cultural forces and the sorts of factors mentioned in §6.3.1, have exerted an influence its content as well. The evolutionary account, for example, is consistent with Philip Kitcher's proposal that the content of our moral sensibility has been largely shaped by a process of *cultural* evolution (as opposed to evolution by natural selection):

> During at least fifteen thousand years, different lineages of our Paleolithic and Neolithic ancestors explored virtually all the systems of rules and ideals for regulating conduct that have figured in the every day conduct of most people (including most contemporary people). Many of these systems did badly in cultural competition: the groups that adopted them were not very good at transmitting their ideas to contemporaries and descendants. The systems that survived were absorbed in later moral practices and figured in the codes that emerge in the Mesopotamian and Egyptian texts. Cultural evolution continued as

[18] I borrow this example from Ruse (1986/1998: 238-242) and Street (2006: 115). For more examples of moral judgments that appear likely to enhance the reproductive fitness of individuals who tend to make or subscribe to them, see Street's paper.

> the central themes are transmitted to the Hebrews and Greeks (Kitcher 2005: 174).[19]

Note also that the evolutionary account is compatible with the claim that content of our current moral sensibility has been greatly influenced by the kind of coherence reasoning favored by many contemporary ethicists (such as Boyd [1988], Brink [1989], Daniels [1979], Sturgeon [2002], and Rawls [1971/1999]). This is consistent with the evolutionary account insofar as it is allowed that the content of the moral sensibility had by our earliest moralizing ancestors serves as the starting point from which coherence reasoning (thought of as a collective social enterprise) began.

6.3.3. From the Darwinian account to moral anti-realism.

Most importantly, it should be observed that in the evolutionary account sketched above, no mention is made of genuine, stance-independent moral facts (Blackburn 1988: 363; Gibbard 1990: ch. 6; Joyce 2006: ch. 6; Kitcher 2005: 175; Ruse 1986/1998: 250-256; Street 2006: 125-135). According to this account, the capacity for moral judgment did not confer advantages on our ancestors because it allowed them to detect important facts about their environment that were awaiting discovery; rather, the advantages were reaped because of the effect that this capacity had on restraining certain behaviors that were maladaptive or otherwise risky from the point of view of reproductive fitness. The moral sensibilities of our ancestors could fulfill this function (of motivating adaptive behaviors) even if there were no facts that their sensibilities represented. Because this evolutionary explanation of our moral sensibility and its contents does not require that we posit the

[19] Another proponent of cultural evolution as a mechanism for shaping moral norms is Shaun Nichols. In his (2002), he argues that moral (and other) norms that prohibit actions that tend to elicit negative emotions have a higher likelihood of surviving to later generations than those that do not.

existence of stance-independent moral facts, it follows that, if it offers the best available explanation of our moral sensibility, and if we accept EC, then we must accept an anti-realist construal of moral inquiry.

6.3.4. Evidence favoring the Darwinian explanation of moral theory.

Of course, the account I have sketched here is speculative. It is also coarse-grained and incomplete. Why believe that it, or something sufficiently close to it, is correct? A fully satisfactory answer to this question would require a comparison with competing hypotheses about the origin and content of our moral sensibility. Unfortunately, there isn't the space to pursue such a task here, at least with respect to all the competing non-moral explanations of moral sensibility that might be adduced. I will, however, address one realist hypothesis below in §6.4.1. But before turning to that hypothesis, I want to indicate several considerations that reflect favorably on the present hypothesis, even if they fall far short of confirming it.

In the first place, as Richard Joyce observes, the tendency to make moral judgments "exists in all human societies we have ever heard of" and "exists in virtually every human individual," and develops within "virtually every human individual…without formal instruction, with no deliberate effort, and with no conscious awareness of its special features" (2006: 134, 135). According to Joyce, these observations "strongly suggest that the tendency to make moral judgments is innate" (*ibid*. 137). The evolutionary account nicely explains how such an innate tendency might arise. Indeed, it is hard to see how anything other than a Darwinian account *even could*

plausibly explain the innateness of this (or any other) tendency, given a commitment to metaphysical naturalism.[20]

A second consideration favoring the evolutionary hypothesis is this: According to this account, human moral sensibility developed out of the conative states of our pre-moral ancestors. If it should turn out that our own moral judgments are driven largely by emotion, this would seem to favor the evolutionary account over those accounts that suggest that our moral sensibilities developed as a matter of detecting stance-independent facts (Joyce 2006: 128, 130). As it happens, the psychological evidence seems to indicate that emotion is the driving force behind moral judgment (see, for example, Haidt 2001).

Finally, the evidence from primatology seems to favor the evolutionary story. Present-day chimpanzees exhibit an impressive array of altruistic and cooperative behaviors (de Waal 1982). This fact gives us some reason to expect that our non-moralizing hominid ancestors had affective dispositions that produced altruistic behavior. If so, it is hard to see why we should doubt that these dispositions have survived in us and have exerted an influence on our present-day moral and evaluative judgments.

I believe that the considerations mentioned above give us reason to conclude that the evolutionary account, or something relevantly like it, is (at the very least) a contender for the title of best available explanation of our current moral sensibility (at least, when this account is supplemented with a description of further cultural influences, including

[20] I can think of only two competing hypotheses. The first hypothesis is that God implants these tendencies in our minds at birth. The second is something like a Platonic theory of recollection: these tendencies were acquired in past lives and are merely "recollected" by each living human. Setting aside the question of whether these hypotheses are plausible (it is not at all clear how one could recollect a *tendency*, as opposed to a proposition), neither seem compatible with naturalism, at least granting widely held assumptions that many naturalists accept (*viz*., that neither God nor past lives are needed in the best explanations of our observations).

our engaging in coherence moral reasoning). It should be noted that, because it does not require that we posit the existence of moral facts, this evolutionary explanation is very likely to be more parsimonious than any explanation of our moral sensibility that would be more congenial to moral realism. Consequently, even if the realist is able to produce an account that is equally plausible, the evolutionary story would retain a theoretical advantage since it requires us to posit fewer kinds of entities than a realism-friendly explanation would require. In that case, the best explanation of our moral sensibility would not require that we posit the existence of stance-independent moral facts; and so, in accordance with EC, we should reject moral realism.

6.4. Return of the Tu Quoque?

6.4.1. Tracking accounts of moral sensibility.

How might a moral realist reply to this revised argument from explanatory impotence? He could reject EC, of course. But as we saw in §6.1, this maneuver is not available to ethical naturalists.[21] Another line of reply would deny that the evolutionary account sketched above offers the best available explanation of the origin and content of our moral sensibility. For this strategy to be persuasive, the realist needs to offer his own competing account of the development of our moral sensibility. Such a story must have it that our ancestors' moral sensibilities (or proto-moral sensibilities) were shaped by and

[21] But even if a realist rejects EC and accepts ethical non-naturalism, it is far from clear that the threat posed by the evolutionary account of moral sensibility has been defused. The non-naturalist would still owe an account of the epistemic processes or mechanisms by which we transcend the influences of evolution on our moral sensibilities in order to grasp the non-natural moral facts. The non-naturalist cannot simply allow that our present moral knowledge is merely the result of our subjecting our innate moral sensibility to coherence reasoning. For, in that case, we would have no reason to think that our moral beliefs accurately represent the non-natural facts. Thus, while the evolutionary account of moral inquiry is not *per se* incompatible with non-naturalistic realism, it certainly seems to encourage moral skepticism.

responsive to stance-independent moral facts. Following Sharon Street (2006), we can call accounts of this sort *tracking accounts* of moral sensibility.

There are at least two salient forms that tracking accounts of moral sensibility might take. In one form, it is conceded that humans and their ancestors have innate moral sensibilities. Such an account, presumably, would take the form of an alternative evolutionary story. According to this kind of story, the ability to detect moral facts conferred a reproductive advantage on our ancestors and, thus, developed as a result of natural selection. A tracking account need not suppose that humans evolved an innate moral sensibility, however. One might argue, instead, that we are able to detect—and, thus, track—moral facts using the same cognitive equipment that we use to detect non-moral facts. According to this kind of tracking account, we are able to perceive moral facts either through direct observation, or else by abductively inferring them from our observations.

With respect to tracking accounts of the latter sort, I take it that there is little hope that our moral sensibilities could have arisen by way of direct perception of moral facts.[22] The more promising question to pursue is whether we might have arrived at our current moral sensibility by way of abductive inference from our observations of non-moral facts. To ask this, in essence, is to ask whether moral facts are needed in the best explanations of observable phenomena—where the phenomena in question are something other than the mere fact that we have a moral sensibility.[23] I will consider whether there

[22] Even if we allow that there are cases of moral perception, as Sturgeon notes, those perceptions would depend upon the existence of a background moral theory (perhaps tacitly) held by the agent (1985a; Boyd 1988). But in that case, the perceptions of moral facts cannot be the explanation for the development of our moral sensibility; the existence of the sensibility is ontologically prior to our supposed moral perceptions.
[23] The reason for this exclusion is that the very purpose of our search for explanandum is in order to show that moral facts figure in the best explanation of our sensibility. If we include the existence and content of our sensibility among the possible explanandum, the tracking account that emerges will be circular. The

is any serious abductive work for moral facts to do in Chapter 7. My discussion there centers on the question of whether moral theories yield interesting empirical predictions; nevertheless, the very same putative predictions double as examples of phenomena that are potentially best explained by moral facts.

Let us turn our attention, then, to the innate sensibility version of the tracking account. The trouble for this kind of account is that there does not yet exist a plausible story of how the capacity to detect stance-independent moral facts (and to recognize them *as* moral facts)[24] might have conferred a reproductive advantage on organisms. Nor is it easy to see how an illuminating story of this kind might go. But even if we were confident that one could be given, it must be noted that it would have to be noticeably superior in certain respects to the anti-realist's account sketched in §6.3. As we saw, the anti-realist's explanation does not require that we posit the existence of moral facts. Because of this, it is more parsimonious than any tracking account of moral sensibility that might be offered. Consequently, a viable tracking account will need to be superior to the anti-realist account in some other respect, if it is to lay claim to the title of the best explanation of our moral sensibility.[25]

---

account would say, in essence, that the evidence for holding that moral facts are needed to best explain our moral sensibility is the fact that we need to posit such facts in the best explanation of our sensibility.

[24] This qualification is needed because, again, some naturalists identify moral properties with natural properties such as *maximizing the balance of pleasure over pain*. I contend that it is not enough that a tracking account provides some story about how the ability to detect whether something maximizes the balance of pleasure over pain yields a reproductive advantage. A creature with no moral sensibility whatsoever could be a flawless detector of this kind of fact. Because of this, it should be clear that an explanation merely of how we detect natural properties of this sort will not suffice to explain how our ancestors developed their moral sensibility. What is needed is a story about how the supposed fact that a natural property of this sort is identical with *moral rightness* played a role in the evolution of our supposed ability to recognize this property "under the guise" of *rightness*. The question, then, is this: "how and why did the ability to see a property such as *maximizing the balance of pleasure over pain* as right-making increase the reproductive fitness of creatures with this ability?"

[25] Street makes the same point in her (2006: 129)

In light of the high bar of theoretical success that any tracking account of moral sensibility must meet, it seems that, while we await the development of such an account, we are justified in being pessimistic about its prospects for success. Let us turn our attention, then, to a different kind of response to the revised argument from explanatory impotence.

6.4.2. Debunking explanations of scientific thought.

As we saw in §6.2.3, Harman's original argument from explanatory impotence is vulnerable to a *tu quoque* reply. I have suggested that the revised argument from explanatory impotence, which incorporates an evolutionary explanation of moral sensibility, is not vulnerable to this sort of reply on behalf of moral realism: whereas the revised explanatory impotence argument shows that moral facts are not needed in the best *a posteriori* explanations of our moral sensibility, a similar line of reasoning cannot be used to show that scientific or theoretical facts (e.g., facts about the existence and nature of theoretical entities such as protons, electrons, quarks, etc.) are not needed in our best *a posteriori* explanations of our accepting the scientific theories that we accept. But perhaps this is not so. Perhaps the same kind of argument could be employed in the service of scientific anti-realism. If it could, this would show that the argument from explanatory impotence cannot be wielded against moral realism by philosophers such as Harman and me, who accept scientific realism (i.e., realism about the sorts of facts and entities posited by our best scientific theories).

6.4.3. Railton's evolutionary *tu quoque*.

In response to evolutionary explanatory impotence arguments of the kind that I sketched in §6.3, Peter Railton writes:

> …[T]his seemingly hard-headed argument is really more of a threat to itself than to morality. For it presupposes a normative premise that it tends by its own reasoning to undercut. Why is it 'facing facts' to force moralists to confront theories of natural selection? – Because these theories are epistemically well confirmed. Who confirmed them, and how? – Humans did, by using scientific methods. But this assumes that humans are psychologically and socially equipped to carry out scientific inquiry, to produce and test hypotheses in ways that yield impartial epistemic justification, despite the fact that our perceptual, cognitive, linguistic, and deliberative capacities have all been shaped by a process of natural selection in which opportunism—not impartiality, warrant or truth—rules. Why, then, isn't human epistemic pretense illusory? The hard-headed argument hammers itself into the same ground into which it had previously pounded morality (2000: 57).

This counter to the evolutionary explanatory impotence argument fails. There is no doubt that Railton is correct that our perceptual, cognitive, linguistic, and deliberative capacities have been shaped by natural selection. The trouble is that, unlike the case of moral thought, there is simply no plausible story to tell about the evolution of these capacities according to which they enhanced reproductive fitness without delivering (or at least facilitating) a roughly accurate representation of reality to the organisms that have them. This is most obvious in the case of perception. The capacity to experience perceptual representations of a prey animal in location L won't enhance a predator's reproductive fitness much unless this perception regularly correlates with a prey's actually being in L. This example suggests that the best evolutionary explanation of our perceptual faculties will be a tracking account. Indeed, it is hard to see how any non-tracking explanation could even be plausible.

A similar thing can be said of our basic inferential capacities. Assuming that our basic inferential capacities were formed by a process of natural selection, we need some explanation of how it is that these capacities enhanced the reproductive fitness of our evolutionary ancestors.[26] It seems to me that there is reason for optimism about tracking accounts on this front as well. To begin with, it is at the very least doubtful that a tendency to make mostly false inductive inferences could have *enhanced* reproductive fitness. It is almost certain that such a tendency would instead diminish a creature's fitness. Quine is surely correct when he writes, "Creatures inveterately wrong in their inductions have a pathetic but praiseworthy tendency to die before reproducing for their kind" (Quine 1969: 126). If this is right, then to the extent that natural selection explains our having the inferential tendencies that we have, it is most likely the evolutionary pressure has been directed towards producing largely accurate or cogent inductive inferences.[27]

Richard Joyce, responding to the same passage from Railton, makes a similar point with respect to basic arithmetical beliefs. Suppose that simple arithmetical beliefs, such as 1 + 1 = 2 are innate, and thus, are likely the result of natural selection.[28] Even if

---

[26] It is possible, of course, that our inferential capacities are a "spandrel", a mere by-product of some different trait that was selected for. All I want to suggest below, however, is that, to the extent that a capacity to make inferences is fitness-enhancing, the most likely account will be one in which the degree to which this capacity is fitness-enhancing is positively correlated with the degree to which it produces *accurate* conclusions.

[27] To avoid misunderstanding, note that I am not suggesting that natural selection has produced (or must produce) in us *perfectly* accurate inferential tendencies that are reliable no matter what sort of physical environment we may find ourselves in. Indeed, it is well known that human beings have a number of inferential habits that are unreliable in many contexts. But this should not discourage us from thinking that that our inferences are often fairly reliable—or at any rate, that they are not "inveterately wrong." (For a discussion of human inferential failures, see Nisbett and Ross [1980]).

[28] Although Joyce cites evidence for the view that "natural selection has provided humans with an inbuilt faculty for simple arithmetic," he does not claim that the belief that 1 + 1 =2 in particular is innate. The innateness of this belief is assumed merely for the purposes of illustration (Joyce 2006: 182).

we conclude that these beliefs really have come to us via natural selection, Joyce argues, it does not follow that we have reason to doubt their accuracy:

> …[W]e have no grasp of how this belief might have been selected for, how it might have enhanced reproductive fitness, independent of its truth. False mathematical beliefs just aren't going to be very useful. Suppose you are being chased by three lions, you observe two quit the chase, and you conclude that it is now safe to slow down (Joyce 2006:182).

Joyce does not treat his readers to an ending to this little illustration, but his point is obvious: the creature that fails to conclude that three lions in pursuit minus two leaves one lion still giving chase does not live another day to reproduce. The upshot here is that, if basic mathematical beliefs come to us via natural selection, then any plausible genealogy of those beliefs will be a tracking account. It is hard to see how a tendency to draw largely false mathematical beliefs would enhance the reproductive fitness of organisms.

I conclude then, that in the absence of further argument, the fact that the cognitive capacities grounding our scientific practices have been shaped by natural selection gives us no reason to worry that these capacities fail to deliver a roughly accurate picture of reality. Furthermore, the fact that these capacities have an evolutionary genealogy gives us no reason to worry that the scientific practices we have built on top of them are not truth-conducive.

6.4.4. The social-historical case against scientific realism.

Moral realists looking for debunking arguments against scientific realism to use in a *tu quoque* reply to the moral anti-realist revised argument from explanatory impotence might expect more success by co-opting the arguments that scientific anti-realists have

themselves offered. Recall that the revised argument from explanatory impotence rests on the fact that the best explanation of how we arrived at our current moral sensibility is one that does not require the positing of stance-independent moral facts; the forces shaping our moral sensibility do not in any obvious way depend upon—nor do they seem to be responsive to—supposed moral facts themselves. Some opponents of scientific realism have made a similar charge about our current scientific theories: the best explanations of how we arrived at our present day scientific theories do not require that we posit the existence of mind-independent theoretical facts and entities. Because of this, it is argued that we are unjustified in believing that our current scientific theories are stance-independently true.

The scientific anti-realist's argument begins with the claim that the choice of which theory to accept from a range of alternatives is underdetermined by our observational evidence. For any theory T^1 that accurately accounts for the observable data, it is possible to construct another theory T^2 that accounts for the same data but avoids a commitment to the unobservable entities posited by T^1. Because of this, our acceptance of a given theory out of a range of possible alternatives must always rest on something beyond that theory's mere success in conforming to the observable data. The anti-realist argues that the sorts of factors that lead us to accept of a given theory over its "empirically equivalent" rivals—that is, over rival theories that issue the same predictions about observable phenomena[29]—are epistemically irrelevant: they fail to justify us in thinking that our preferred theory, rather than its empirically equivalent rivals, is stance-independently true. In this vein, Thomas Kuhn writes,

[29] I draw this definition of 'empirical equivalence' from Boyd (1983: 46; cf. Boyd 1982: 618).

> Observation and experience can and must drastically restrict the range of admissible scientific belief, else there would be no science. But they cannot alone determine a particular body of such belief. An apparently arbitrary element, compounded of personal and historical accident, is always a formative ingredient of the beliefs espoused by a given scientific community at a given time." (1962/1996: 4).

What sorts of arbitrary and accidental factors might have influenced the acceptance of our scientific theories? In the most extreme formulation of this kind of argument, the anti-realist claims that the major factors determining theory acceptance are political, ideological, and/or personal. Convincing examples in which some area of scientific research and its conclusions is driven by these considerations are not hard to find. It hardly comes as a surprise to us that environmental research funded by petroleum companies or right-wing think tanks tend to conclude either that global warming is not anthropogenic, or else that it is not as serious a threat to the interests of mankind as other scientists claim. It is plausible to suppose that these conclusions have more to do with the financial interests of petroleum companies and the political ideology of right-wing think tanks than they have to do with stance-independent facts about the real trajectory and causes of global warming. (Or perhaps this example gets it backwards. Some on the right contend that environmental research in the academy is driven less by the quest for truth than by the desire of left-wing academics to provide a rationale against unfettered capitalism.)[30] Another putative example of this phenomenon is provided by evolutionary psychology. Critics argue that the conclusions arrived at by evolutionary psychologists—namely, claims to the effect that certain psychological traits are innate and the result of natural selection—are little more than an attempt to justify existing power structures.

[30] Thus, Ayn Rand writes, "The immediate goal [of environmentalists] is obvious: the destruction of the remnants of capitalism in today's mixed economy, and the establishment of a global dictatorship" (Rand 1971/1999: 280).

Along these lines, the Sociobiology Study Group writes, "It is not surprising that the model of society that turns out to be 'natural' bears a remarkable resemblance to the institutions of modern market society, since the theorists who produce these models are themselves privileged members of just such a society" (1977: 133).

In general, the greater the role that ideological considerations of these sorts play in the best explanations of our collective acceptance of current scientific theories, the less of a need there is to posit the approximate truth of these theories in order to explain our acceptance.

Other putatively non-epistemic factors that influence the acceptance of scientific theories make for a somewhat less cynical case against scientific realism. Among the factors that determine theory acceptance, according to Kuhn, are "aesthetic" (or "pragmatic") considerations. For example, we change our allegiance from one theory to another because we judge the new theory to be "simpler" or more "elegant." Citing factors of this sort is especially needed to explain theory acceptance in cases where a scientist must choose between two or more empirically equivalent theories. The trouble that this phenomenon poses for the scientific realist is that, according to his anti-realist opponents, the aesthetic or pragmatic features of a theory are of no epistemic significance:

> In so far as they go beyond consistency, empirical adequacy, and empirical strength, [pragmatic considerations] do not concern the relation between the theory and the world, but rather the use and usefulness of the theory; they provide reasons to prefer the theory independently of questions of truth (van Fraassen 1980: 88).

More than this, judgments about which theory is simpler or more elegant are sometimes held to be subjective, admitting of no single correct standard. Because choices about

which theory to accept are made by appeal to these subjective aesthetic standards, there are no grounds for saying that one scientist's favorite theory has a greater claim to being correct than another scientist's favorite theory, at least where those theories are empirically equivalent. Because of this, Kuhn writes, "There is no neutral algorithm for theory choice, no systematic decision procedure which, properly applied, must lead each individual in the group to the same decision" (1969/1996: 200, cf. 184ff).

For the scientific anti-realist of this stripe, then, the best explanation of why we currently accept the scientific theories that we do does not require that we understand those theories to be approximately stance-independently true. The best explanation of why we accept a theory T^1 over its empirically equivalent rivals must cite primarily aesthetic, personal, or political factors. However, these factors could motivate our adoption of T^1 even while some other theory T^2 is true (where T^2 has different ontological commitments from T^1). Because of this, have no reason to suppose that the entities posited by T^1 are part of the best explanation of our accepting T^1. Thus, in light of EC, we are not justified in positing the existence of the theoretical entities that our current scientific theories apparently commit us to (e.g., protons, quarks, electromagnetic fields, etc.). If this kind of argument is no less compelling than the argument of §6.3, then we must conclude that the revised explanatory impotence argument against moral realism fails to avoid the realist's *tu quoque* complaint. Either the moral anti-realist must also reject scientific realism—which he is loath to do—or else he must concede that the revised explanatory impotence argument against moral realism is a failure.

6.5. Breaking the Tu Quoque: The Case for Scientific Realism.

6.5.1. Overview.

Are the anti-realists' explanations of our current scientific theory really the best available? According to one standard argument in favor of scientific realism, they are not: anti-realists cannot provide satisfactory explanations of the "success" of scientific theories and scientific practice. The scientific realist contends that the best explanation for the success of science is that our current scientific theories are approximately true and that the putative entities to which theoretical terms putatively refer really exist. This argument—the so-called "ultimate argument" for scientific realism—has been advanced, in one form or another, by Boyd (1982), Kitcher (2001), Jarrett Leplin (1997), Alan Musgrave (1988), Hilary Putnam (1975c), and J.J.C. Smart (1963), among others. If successful, this argument would block the *tu quoque* reply to the revised argument from explanatory impotence against moral realism. In §6.5.2 and §6.5.3 I present what could be called the standard version of this argument for scientific realism. In §6.5.4 I go on to present Boyd's own particular version the argument. In §6.5.5 I consider the prospects for the revised explanatory impotence argument against moral realism, should all forms of the ultimate argument for scientific realism fail.

6.5.2. The standard case for scientific realism.

Scientific realism can be understood roughly as the view that there are theoretical facts (e.g., facts about unobservable, theoretical entities posited by scientific theories) and that these facts obtain independently of the beliefs and theories of scientists (and others). Boyd adds to this characterization the further claim that "scientific theories should be

understood as putative descriptions of" these theory-independent facts (1988: 181). Assuming that our current physical theories are (approximately) true, we should understand the scientific realist to be committed to the claim that there exist things such as protons, neutrons, and electrons, and that the existence of these things—and the properties that they have—depend neither on our believing that they exist, nor on the fact that we would believe that they exist were we in ideal epistemic conditions.

The standard version of the ultimate argument for scientific realism begins with the premise that our current scientific theories are instrumentally reliable to a high degree. As Boyd characterizes it, the *instrumental reliability* of a theory is a measure of "its ability to provide…approximately accurate predictions about the behavior of observable phenomena" (1982: 616). For example, the theory of relativity is instrumentally reliable to a certain extent in virtue of its ability to predict accurately the deflection of light passing by the Sun and to predict accurately the perihelion precession of Mercury (Leplin 1997: 78-80; Will 1986: chh. 3, 4). Several philosophers who advance the ultimate argument emphasize the importance of "novel" predictions when it comes to assessing the instrumental reliability of a theory. Following Musgrave, we can say, roughly, that "a predicted fact is a novel fact for a theory if it was not used to construct that theory—where a fact is used to construct a theory if it figures in the premises from which that theory was deduced" (Musgrave 1988: 232; cf. Boyd 1983: 54; Leplin 1997: 77).[31] By this characterization, the theory of relativity's prediction of the Sun's deflection of light and Mercury's perihelion precession count as novel.[32] By

[31] I would also include Paul Thagard's discussion of "conservative dynamic consilience" as an endorsement of novel prediction as an important element of theory confirmation (Thagard 1978: 83f).
[32] As Leplin characterizes novel prediction, the Sun's deflection counts as novel for the theory of relativity, but the precession of Mercury's perihelion does not. He takes the latter prediction to be non-novel because

contrast, the predictions of the Ptolemaic model of the solar system concerning the positions of the Sun, Moon, and planets do not count as novel; for the Ptolemaic theory was constructed precisely in order to fit the known data concerning the recurring positions of such bodies.[33] Consequently, to the extent that these latter sorts of predictions allow us to say that the Ptolemaic theory is instrumentally reliable at all, we should say they confer on it only a modest degree of instrumental reliability.

The more controversial premise figuring in the scientific realists' ultimate argument is this: the best—indeed, the only credible—explanation for the high degree of instrumental reliability exhibited by our current scientific theories is that such theories are approximately (stance-independently) true. More specifically, the claim is that the best (and perhaps the only plausible) explanation of the instrumental reliability of our scientific theories is that the central theoretical terms utilized by those theories successfully refer to stance-independent theoretical entities, and that these entities by and large have the properties that our theories ascribe to them. In short, in order to best explain the instrumental reliability of our current scientific theories, we must construe these theories realistically. Along these lines, scientific realists write:

> If the phenomenalist [i.e., anti-realist] about theoretical entities is correct we must believe in a *cosmic coincidence*. That is, if this is so, statements about electrons, etc., are of only instrumental value: they simply enable us to predict phenomena

---

it fails his "uniqueness condition," which requires that no existing plausible alternative theory predicts a qualitatively similar result. Because Newtonian theory also predicted a precession of Mercury's perihelion—albeit with less quantitative accuracy than relativity does—the prediction of the precession by the theory of relativity is not unique, and so not novel according to Leplin's account. By contrast, although Newtonian theory could also be used to predict the bending of light, it required discredited auxiliary hypotheses about the nature of light to do so. For this reason, Leplin claims that relativity's prediction of the bending of light passing by the Sun satisfies his uniqueness condition for novelty (Leplin 1997: 77-80).

[33] To ward off objections, I recommend that we read Musgrave's criterion of novelty as concerning fact-*types*, rather than fact-*tokens*. Because every prediction is directed towards hitherto unobserved future token-events, there is a trivial sense in which no interesting fact that we could care to predict figures in the premises of a putatively empirical theory. (In this vein, a defender of Ptolemy could object that her successful predictions about an eclipse that is to occur in 2024 ought to count as novel, since Ptolemaic astronomers could not have appealed to this token-fact in the construction of their theory.)

> on the level of galvanometers and cloud chambers. They do nothing to remove the *surprising character* of these phenomena. […] On the other hand, if we interpret a theory in a realist way, then we have no need for such a cosmic coincidence: it is not surprising that galvanometers and cloud chambers behave in the sort of way they do, for if there really are electrons, etc., this is just what we should expect (Smart 1963: 39).

> The positive argument for [scientific] realism is that it is the only philosophy that doesn't make the success of science a miracle. That terms in mature scientific theories typically refer…, that the theories accepted in a mature science are typically approximately true, that the same term can refer to the same thing even when it occurs in different theories – these statements are viewed by the scientific realist not as necessary truths but as part of the only scientific explanation of the success of science, and hence as part of any adequate scientific description of science and its relations to its objects (Putnam 1975c: 73).

> [The ultimate argument for scientific realism] is, I suggest, an inference to the best explanation. The fact to be explained is the (novel) predictive success of science. And the claim is that realism (more precisely the conjecture that the realist aim for science has actually been achieved) *explains* this fact, explains it *satisfactorily*, and explains it *better* than any non-realist philosophy of science. And the conclusion is that it is reasonable to accept scientific realism…as true (Musgrave 1988: 239, emphasis in the original).

I find the scientific realist's argument persuasive. But there are at least two concerns that deserve to be addressed. The first concerns the notion of "approximate" truth. The second concerns the extent to which the standard version of the ultimate argument addresses the scientific anti-realist's objections surrounding the use of non-epistemic considerations when deciding between theories. I will address these concerns in the next two sections.

6.5.3. Approximate truth.

If the ultimate argument is to be plausible, it must be formulated utilizing the notion of approximate truth rather than *precise* truth, or truth *simpliciter*. Boyd writes that "No realist conception that does not treat theoretical knowledge and theoretical progress as

involving approximations to the truth is even prima facie compatible with the actual history of science" (1990: 216). Perhaps what Boyd has in mind is the sort of worry described by Thomas Weston:

> The history of science provides abundant evidence for what Newton-Smith calls the "dismal induction": Theories which have gained and deserved acceptance have almost always turned out to be false. Past experience would indicate that even if a theory is sufficiently supported to warrant acceptance, the probability that it is precisely true is roughly zero (Weston 1992: 54).

In addition, Leplin, who cashes out the notion of approximate truth in terms of "accuracy of representation," notes that "there is no reason to suppose that complete or unimprovable accuracy of representation is required for the explanatory or predictive adequacy of the mechanism that a theory postulates" (1997: 103). If he is correct, then a scientific realist could not claim that the only plausible explanation for the instrumental reliability of our current scientific theories is that these theories are precisely true. Such a claim would be much stronger than the evidence warrants. Thus, the key premise (and the conclusion) of the ultimate argument must be formulated in terms of approximate—rather than precise—truth.

The ultimate argument for scientific realism cannot succeed, then, unless we can make sense of the notion of approximate truth. Unfortunately, the notion of approximate truth is both controversial and difficult to formalize. Here are several attempts at a rough account of approximate truth:

> [A sentence] *P* is approximately true at [world] *u* if and only if *P* is true in some world similar to (or close to) *u* (Hilpinen 1976: 24)[34]

> …[T]o talk of respects of approximation to the truth is to talk of respects of similarity and difference between actual causal situations and certain possible ones (Boyd 1990: 239).

[34] In the original, Hilpinen italicizes the entire sentence. I have removed those italics.

> The basic idea of the definition of approximate truth is that a statement will be approximately true under interpretation *I* if there is an interpretation *J* which is "near" *I*, and under which it is actually true (Weston 1992: 60f).
>
> A realist interpretation attributes some measure of truth to the theory, where truth is understood as accuracy of representation. As accuracy comes in degrees, it is natural to speak of "partial" or "approximate" truth, or of truth in "some measure" (Leplin 1997: 103)

Roughly, we may say that a theory is approximately true just in case the processes and mechanisms that it posits and describes are "sufficiently similar" to actual processes and mechanisms in the world. This is not the place to attempt a defense or formalization of approximate truth (for the latter, see Hilpinen [1976] and Weston [1987]). Fortunately, in the context of the current dialectic, a defense is not needed. We have already seen that Boyd is committed to the legitimacy of approximate truth in his defense of scientific realism. More importantly, both he and the rest of the principal defenders of synthetic ethical naturalism make use of the notion of approximate truth in their defenses of moral realism (Boyd 1988: 201, 207, 209; Brink 1984: 24; 1989: 129, 299; 1999: 207; Sturgeon 1985a: 67f, 72; 1986a: 73f; 1992: 99, 108; 2006a: 254f). In the present context, then, the legitimacy of approximate truth is not in question. If approximate truth should turn out to be an unworkable notion, then so much the worse for the naturalist's case for moral realism.

6.5.4. Non-epistemic methodological principles and Boyd's version of the ultimate argument.

Boyd agrees the standard version of the ultimate argument for scientific realism succeeds in showing that there is something wrong with the denial of scientific realism. He observes, however, that the argument does not offer a direct rebuttal to the scientific anti-

realist's argument that our reliance on non-epistemic methodological principles in deciding between empirically equivalent theories implies that we are unwarranted in supposing our preferred theories are approximately true. A satisfactory version of the ultimate argument should address this claim (Boyd 1983: 54).

In advancing his own version of the ultimate argument for scientific realism, Boyd contends that it is not merely the instrumental reliability of scientific *theories* that needs explaining, but also the instrumental reliability of scientific *methods*. For Boyd, a collection of methodological principles is instrumentally reliable to the degree that the implementation of these methods tends to lead scientists to accept instrumentally reliable theories (Boyd 1982: 16; 1983: 76; 1990: 221). As examples of instrumentally reliable methodological principles, he offers the following:

> (1) *Conservatism*: "…new theories should, *prima facie*, resemble current theories with respect to their accounts of causal relations among theoretical entities" (1973: 7; cf. 1982: 618f).
>
> (2) *Principle of experiment design*: "…a proposed theory *T* must be experimentally tested under situations representative of those in which, in the light of collateral information, it is most likely that *T* will fail, if it is going to fail at all" (1973: 10; cf. 1982: 629f).
>
> (3) *Principle of measurement procedures* "…one should follow the dictates of the best confirmed theory in (re)designing measurement procedures" (1983: 79; cf. 1982: 19f).

Boyd notes that all three of these principles are "theory-dependent." This is obvious in the case of conservatism: whether or not a new theory is acceptable from the point of view of conservatism depends upon what theory (or theories) we presently accept. But even the principle of experiment design is theory dependent: judgments about the sorts of conditions under which a new theory is likely to fail must be made on the basis of our current scientific background knowledge, which may include propositions arrived at on

the basis of the current theory we are looking to replace. Finally, the principle of measurement procedures explicitly directs us to consider the theories we presently accept when designing measurement procedures.

According to Boyd, the fact that methodological principles that are so theory-dependent are nevertheless instrumentally reliable—the fact that they continually lead us to accept increasingly instrumentally reliable theories—stands in need of explanation (1973: 3). He argues that the only scientifically plausible explanations for the instrumental reliability of methodological principles like those mentioned above require us to judge that our scientific theories are approximately (stance-independently) true. According to Boyd, it is because our background scientific theories are approximately true that the implementation of these theory-dependent methodological principles is able to guide us towards accepting increasingly instrumentally reliable theories. Boyd contends that no other hypothesis can make adequate sense of the instrumental reliability of these methods (Boyd 1973: 11f; 1982: 621f; 1983: 64ff).

If Boyd is correct, then an important consequence follows: because our theory-dependent methodological principles are operating on approximately true background theories, there is reason to expect that the application of these principles will be reliable in guiding us towards new theories that are themselves approximately true. In short, the application of these methodological principles is part of a reliable process of true belief production. As a result, such methodological principles have epistemic import after all, contrary to what anti-realists like Kuhn and van Fraassen have claimed: that a new theory

has been arrived at and endorsed through the application of these principles is evidence that the theory is approximately true (Boyd 1982: 622f; 1983: 67).[35]

Naturally, Boyd's argument promises to explain only why aesthetic considerations or seemingly pragmatic considerations such as parsimony and elegance are capable of contributing to the epistemic warrant of scientists' acceptance of certain theories. Since it is doubtful that ideological considerations for selecting theories are likely to be truth-conducive, Boyd's argument may do little to confer epistemic value upon those kinds of considerations. But this is hardly cause for alarm. To the extent that our current theories are highly instrumentally reliable, it is doubtful that a credible case can be made that mere ideological considerations have been the most significant grounds upon which such theories are accepted by scientists. It is surely too incredible to be believed that the principle "select the scientific theory that whose acceptance would most benefit the interests of capitalists over workers" would tend to produce instrumentally reliable scientific theories in the long run (especially in the natural sciences). More likely, instrumentally reliable scientific theories are arrived at in spite of—rather than because of—the influence of political ideology.

6.5.5. <u>What if the ultimate argument is a failure?</u>

As with the standard version of the ultimate argument for scientific realism, I find Boyd's argument to be persuasive. If his argument is indeed successful, then there would seem to be little reason to worry that the kind of reasoning behind the revised explanatory impotence argument against moral realism could be used to upset our commitment to

[35] This inference should be uncontroversial if we follow Boyd in assuming that a reliabilist account of epistemic warrant is correct. Again, since Boyd and his fellow ethical naturalists *do* accept epistemological reliabilism in one form or another, this assumption is safe in the context of the present dialectic.

scientific realism: the best explanation of our accepting the scientific theories that we accept is a realist explanation. Consequently, moral anti-realists and ethical non-naturalists can advance the explanatory impotence argument against naturalistic moral realism without fear of endangering their commitment to scientific realism, if they have such a commitment.

As I said, I am inclined to think that the ultimate argument (in at least one of its forms) is sound. But it must be acknowledged the ultimate argument for scientific realism has not produced consensus among philosophers of science. Scientific anti-realists point to numerous cases in the history of science where a theory proved to be instrumentally reliable despite the fact that its central terms failed to refer, and thus, the theory was not even approximately true. In addition, they have raised objections to the very notion of approximate truth and its use by realists.[36]

While I continue to find the realist's case to be persuasive despite these objections, I must be frank and admit that I do not have the scientific expertise to evaluate the anti-realist's replies to either form of the ultimate argument. In place of a further defense of scientific realism, I want to consider what the fallout would be for the explanatory impotence argument against moral realism, if it were it to turn out that the ultimate argument for scientific realism is a failure.

Boyd suggests that the ultimate argument probably "reconstructs the reason why most scientific realists are realists" (1983: 54). I believe that this is true in my own case: it strikes me as very unlikely, for instance, that scientists could have developed an atomic bomb if the physical theory that they were working with was not approximately correct—at least, correct to the extent that the world really contains entities that have a legitimate

[36] Both of these lines of objection to scientific realism are advanced by Laudan (1981).

claim to being the referents of 'proton,' 'neutron,' and 'electron.' In that respect, I am convinced that the best explanation of how mid-20th century physicists were able to produce an atom bomb requires that we posit the existence of protons, neutrons, and electrons. In short, I believe that we must suppose that the prevailing physics of the time was at least approximately true in a stance-independent way.

Now, in light of the importance of the ultimate argument (even in unarticulated and inchoate forms) in persuading philosophers and laypersons of scientific realism, it seems to me that, if the ultimate argument could be shown to be a failure, then it would be perfectly reasonable to give up on scientific realism. But in that case, the moral realist's deployment of the *tu quoque* carries no force: if no version of the ultimate argument for scientific realism succeeds, then those of us who accept EC really ought to reject scientific realism along with moral realism. If scientific anti-realism is the price that a metaphysical naturalist must pay for advancing the explanatory impotence argument against moral realism, then, assuming the failure of the ultimate argument, it is a price that he would have had to pay anyway. To conclude this section: even if no version of the ultimate argument for scientific realism succeeds, the moral realist's *tu quoque* reply still fails as a reply to the argument from explanatory impotence.

6.6. Conclusion.

Harman's original argument from explanatory impotence against moral realism is vulnerable to a quick *tu quoque* reply: just as moral facts are not needed to explain our making this or that moral judgment in response to observing an action, theoretical facts are not needed to explain a physicist's theoretical judgments in response to observable

phenomena. I have suggested that a better argument from explanatory impotence would focus on the explanations of the origin and content of our collective moral sensibility as a whole. I have argued that the most promising explanation is a Darwinian account. Since the Darwinian account makes no essential reference to genuine stance-independent moral facts, and since moral facts do not seem to potentially explain any other phenomena, metaphysical naturalists ought to deny the existence of moral facts.

I have argued, in addition, that this revised argument from explanatory impotence is not vulnerable to the sort of *tu quoque* reply that undoes Harman's original version of the argument. Because of the high degree of instrumental reliability of our best scientific theories and methods, there is no compelling explanation of the origin and content of our acceptance of those theories that does not recognize their approximate truth.

In the next chapter, I will consider whether moral realists can mount a counterattack against the revised argument from explanatory impotence that is modeled on the defense of scientific realism that I sketched in §6.5. According to this line of thinking, there may yet be phenomena that moral facts are needed to explain: namely, the instrumental reliability of moral theories. I will argue that this sort of reply on behalf of moral realism fails.

CHAPTER 7

THE PROSPECTS FOR AN ULTIMATE ARGUMENT FOR MORAL REALISM

7.1. Introduction.

Although the failure of the ultimate argument for scientific realism would do no favors for naturalist moral realists, my own view, again, is that the ultimate argument (in at least some of its versions) is sound. Thus, I believe that a debunking argument of the kind of advanced against moral realism in §6.3 fails when it is directed against scientific realism. But even if we grant the success of the argument for scientific realism, the moral realist has one more hand to play. As we saw, the central argument in favor of scientific realism rests on two claims: (i) our best present-day scientific theories (and methods) exhibit a high degree of instrumental reliability, and (ii) the best explanation of this instrumental reliability requires that we construe those theories realistically. Perhaps the moral realist could utilize a similar argument in order to respond to the argument from explanatory impotence. An argument of this sort would begin with the premise that our best current moral theories are instrumentally reliable to a high degree. To this the realist would add the claim that the best available explanation for the instrumental reliability of such theories requires that we construe them realistically.[1] Let us call this *the ultimate moral argument*. In this chapter, I aim to refute the ultimate moral argument—or, to state my goals more modestly, I aim to raise a good amount of doubt about the prospects for the success of such an argument.

[1] To keep things simple, I am going to set aside discussions of claims to the effect that the methodological principles underwriting moral inquiry (as opposed to particular moral theories) are themselves instrumentally reliable. From my discussion of the putative instrumental reliability of moral theories below, my hope is that it will be evident than such methodological principles—at least in their application to moral matters—do not exhibit an impressive degree of instrumental reliability.

7.2. Moral Theories and Empirical Predictions.

7.2.1. On the instrumental reliability of moral theories.

Recall from §6.5.2 that a theory is instrumentally reliable to the extent that it yields "approximately accurate predictions about the behavior of observable phenomena." Recall also that a theory's ability to yield accurate *novel* predictions carries an especially heavy weight in determining the degree of instrumental reliability that it enjoys. Our present task is to examine whether moral theories exhibit a significant degree of instrumental reliability. We must search, then, for observable phenomena that moral theories might help us to predict. On the face of it, this might seem like a fool's errand: moral theories do not aim to tell us what *is* the case or what *will be* the case; rather, they aim to tell us what *ought to be* the case. This sort of consideration might lead philosophers, including moral realists, to doubt the propriety of asking of a moral theory that it be instrumentally reliable (cf. Nagel 1986: 144; Quinn 1986; Shafer-Landau 2006). Be that as it may, at this point in the dialectic, moral realists who accept naturalism, and thus, accept EC, do not have the luxury of dismissing the demand that moral theories yield predictions about observable phenomena; they desperately need to find explanatory work for moral facts to do. As it happens, SEN proponents have offered several examples of predictions derived from moral theories. After dealing with a preliminary concern, I will discuss three of these examples below.

7.2.2. A bad argument against the instrumental reliability of moral theories.

Here is an argument against the instrumental reliability of moral theories that moves too quickly. Let us suppose that our best moral theory is hedonistic act-utilitarianism (AUh). Consider a case in which an agent, Ann, has several alternative actions open to her. Suppose, further, that only one of Ann's alternatives maximizes hedonic utility. Given our assumptions so far, this alternative is morally obligatory according to our best moral theory. Let us now ask what sort of event AUh predicts will occur in this case: will Ann choose to perform the morally right action or not?

It should be obvious that, in the absence of further premises, AUh yields no predictions about which action Ann will perform. In the face of this fact, one might be tempted to conclude that AUh is not at all instrumentally reliable. It may be, of course, that this failure indicates only that AUh is itself a flawed theory. However, it is easy to see that matters are no different for any rival theory of morally right action: no such theory by itself yields empirical predictions about how any agent will behave. Since it is hard to see what other kinds of facts such a theory might predict besides the actions of agents, it would seem that no theory of morally right action is instrumentally reliable. If this is really the case, then the ultimate moral argument is unsound. (At least, this is so insofar as the argument attempts to establish realism about deontic moral properties such as *rightness* and *obligation*. Further argument would be needed to show that evaluative properties such as *goodness* and *badness* share the same fate.)

7.2.3. Three examples of prediction by moral theory.

I said that the argument sketched in the previous section moves too quickly. Here is why: As Brink and Sturgeon observe, theories in general (be they moral theories or non-moral theories) do not yield empirical predictions in isolation. In order to derive a prediction, a theory must be conjoined with "auxiliary premises."[2] Putnam, for example, notes that Newton's theory of universal gravitation issues no empirical predictions without auxiliary premises that give an inventory of what objects there are in space and what additional forces besides gravity are present (Putnam 1974: 255). Both Brink and Sturgeon contend that, if we allow ourselves to utilize auxiliary premises, we can in fact derive empirical predictions from moral theories and principles:

> Candidate moral principles—for example, that an action is wrong just in case there is something else the agent could have done that would have produced a greater net balance of pleasure over pain—lack empirical implications when considered in isolation. But it is easy to derive empirical consequences from them, and thus to test them against experience, if we allow ourselves, as we do in the scientific case, to rely on a background of other assumptions of comparable status. Thus, if we conjoin the act-utilitarian principle just cited with the further view, also untestable in isolation, that it is always wrong deliberately to kill a human being, we can deduce from these two premises together the consequence that deliberately killing a human being always produces a lesser balance of pleasure over pain than some available alternative act; and this claim is one any positivist would have conceded we know, in principle at least, how to test" (Sturgeon 1985a: 51; cf. Brink 1989: 137; Sayre-McCord 1988: 436f).[3]

Presented more formally, Sturgeon's example is this:

> Example A:
>
> A1. For any act, x, x is morally right iff x maximizes hedonic utility (moral theory).

[2] This is insight is commonly credited to Pierre Duhem (1906/1914: 183-188) and Quine (1951: 38f).
[3] Brink appears to hold that the auxiliary premises that conjoin with moral principles to yield empirical predictions must themselves be moral propositions (Brink 1989: 137, 183). I am not so sure that he is right about this. His own example (represented by B below), contains at least one non-moral auxiliary premise. Furthermore, example C below appears to contain only auxiliary premises that are non-moral.

A2. For any act, x, if x is the deliberate killing of a human being, then x is not morally right (auxiliary moral proposition).
A3. For any act, x, if x is the deliberate killing of a human being, then x does not maximize hedonic utility (empirical prediction).

The prediction A3 is entailed by A1 and A2. Now, it may not be quite right to say that A3 is an empirical prediction, since it can't be confirmed by direct observation. Its confirmation would have to proceed, instead, by way of enumerative induction from direct observations of act-tokens of deliberate killings. I propose that we ignore this complication. In any case, it would do just as well for the purposes of the naturalistic moral realist to substitute A3 with the claim that all *observed* acts of deliberately killing a human being fail to maximize hedonic utility.[4]

Another type of example offered by SEN proponents derives predictions from claims about the moral character of agents. This kind of prediction is illustrated most clearly by Brink:

> My moral belief that good people keep their promises when doing so involves great personal sacrifice has no observational consequences when taken in isolation. But when I conjoin it with my independently supported moral belief that Zenobia is a good person (i.e., my evidence not including Zenobia's promise-keeping behavior), I can obtain the observational consequence that Zenobia will keep her promise to Zelda, even though doing so will involve great personal sacrifice on Zenobia's part (Brink 1989: 137).

From this passage, let us construct a second example of a moral prediction:

<u>Example B</u>:

B1. For any person, x, if x is morally good and x has made a promise that requires great personal sacrifice, then x will keep x's promise (moral theory).

[4] A further complication is this: in order to know that a given act-token maximizes hedonic utility, it is not sufficient that we know the hedonic utility of the observed act; we must also know the hedonic utility of all of its alternatives that go unperformed; but this information certainly cannot be acquired by direct observation. I think this fact may well be a serious problem for those who would use example A as evidence in support of an ultimate moral argument. Nevertheless, as with the previous difficulty, I am prepared to set it aside.

B2. Zenobia has made a promise that requires great personal sacrifice (non-moral auxiliary proposition).
B3. Zenobia is morally good (auxiliary moral proposition).
B4. Zenobia will keep her promise (empirical prediction).

In other passages, Brink and Sturgeon suggest that the property *justice* plays a causal role in the world. Sturgeon writes: "A widespread and longstanding assumption about social justice is that it is a stabilizing condition and injustice a destabilizing one, at least under circumstances common enough to be interesting" (1991: 29). Similarly, Brink writes: "We think that political vices (e.g., social injustice) sometimes cause, and so help explain, instability, protest movements, and revolutions; and we think that the political virtues of a society's laws and institutions (e.g., its social justice) can help explain its stability" (1989: 187; cf. Railton 1986: 191f). If they are correct, then our best theory of social justice—assuming it is stance-independently true—ought to facilitate reliable predictions concerning the stability of a given society, at least when that theory is conjoined with this and other auxiliary propositions.

In order to produce an example of a prediction of this sort, we need a sample theory of justice. Because of its familiarity and its plausibility, I will use John Rawls's two principles of justice for my illustration. According to Rawls, the basic structure of society is just if and only if it satisfies the following two principles (with the first given priority over the second):

> [1] Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all. […]
>
> [2] Social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged, consistent with the just savings

principle,[5] and (b) attached to offices and positions open to all under conditions of fair equality of opportunity (Rawls 1971/1999: 266).

I will refer to the conjunction of these principles as 'RPJ' (for 'Rawls's Principles of Justice').

With this theory of justice in hand, I can sketch the third and final example of an empirical prediction generated by a moral theory:

Example C:

C1. For any society, x, x is just if and only if the basic structure of x satisfies RPJ (moral theory).
C2. For any society, x, if x is just, then x is stable (non-moral auxiliary proposition).
C3. The basic structure of society S satisfies RPJ (non-moral auxiliary proposition).
C4. Society S is stable (empirical prediction).

Examples A, B, and C represent the most significant examples of empirical predictions derived from moral theories that are found in the writings of the principal defenders of synthetic ethical naturalism. I believe that, as Brink and Sturgeon contend, these examples succeed in showing that moral theories really do have empirical consequences, at least when conjoined with certain auxiliary assumptions. Thus, there is nothing in principle that prevents a moral theory from being instrumentally reliable to some degree. Unfortunately for ethical naturalists, this finding is not enough to ground a successful ultimate moral argument. In the first place, it is not enough that some moral theories have empirical consequences. What the ultimate argument requires is that our *best* current theories have such consequences, and, moreover, that these consequences are in fact born out by our observations. In other words, our best theory must not only

[5] A just savings principle specifies the kinds and amounts of things that any given generation must save for future generations (Rawls 1971/1999: 252ff). The matter of which savings principle really is just is not important for my purposes here.

predict; it must predict *successfully*. But even this does not go far enough. For it must also be the case that a realist construal of our best theories offers the best explanation of this predictive success. To some extent, whether or not a realist explanation is best will depend upon the character of the predictions and their relationships with the relevant theories. Scientific realists and scientific anti-realists alike acknowledge that some false theories are able to generate reliable predictions. For instance, after it was discovered that metals weigh more after combustion, the phlogiston theory of combustion was amended so as to include the proposition that phlogiston—which was supposed to be released from an object during combustion—had "negative weight" (Thagard 1978: 78; Kuhn 1962/1996: 71). With this amendment, the phlogiston theory would now correctly "predict" of any piece of metal that it would weigh more after it was burned (with the explanation being that, during combustion, the metal releases its phlogiston and hence, loses some of its negative weight). It should be obvious, however, that the best explanation of the amended phlogiston theory's predictive "success" here does not require us to read that theory realistically. (This should be obvious, at the very least, because we now know that phlogiston does not exist.) Indeed, this sort of phenomenon illustrates one reason why scientific realists have held *novel* predictions to be of special importance in confirming a theory. In the phlogiston example, the prediction that metals lose weight upon combustion is clearly not novel to the mature phlogiston theory; the amendment that makes these successful predictions possible would have been completely unmotivated if the relevant data concerning the weight of combusted metals had not already been known.

In the remainder of this chapter, I will examine examples A, B, and C. I will argue that none of these provide a compelling example of a (i) currently best moral theory that (ii) yields successful empirical predictions, where (iii) the best explanation for this success requires a realistic reading of that theory.

7.3. Example A: Predictions Grounded in Two Deontic Moral Principles.

7.3.1. The implausibility of premise A2.

An initial difficulty with example A concerns the auxiliary premise A2. If A is to be a compelling example of a successful moral prediction that supports a realistic construal of a moral theory, then the relevant auxiliary premises used to generate the predictions ought to be plausible. Unfortunately, A2 is not all that plausible. It is surely only the rare pacifist (or perhaps someone who is already in the grip of an explicit moral theory) who thinks that it is in all cases morally wrong to kill another human being. I believe that a more plausible auxiliary proposition would allow that killing is permitted in cases of self-defense. Moreover, those of us who think that not all human beings are persons will not find A2 plausible, even when it is amended to permit self-defense. Perhaps then, the auxiliary hypothesis we are looking for is this:

> A2': For any act, x, if x is the deliberate killing of a person in a non-self defense scenario, then x is not morally right.

While some philosophers might find A2' plausible, others might contend that it is false insofar as it fails to accommodate certain "doomsday" scenarios. For instance, one might think that it is morally permissible to deliberately kill a non-threatening person if that person's sacrifice by us is required in order to avert a catastrophic disaster (e.g., the destruction of 90% of the Earth's human population). One way to deal with this sort of

case is to reformulate A2 as a *ceteris paribus* principle. Another solution, which I will utilize below, is simply to amend it with a "non-doomsday" clause.

In addition to the non-doomsday clause, I would recommend another amendment. As it stands, A2' implies that many—if not most—cases of voluntary active euthanasia are morally wrong. For my own part, I have a firm intuition that many such cases are not morally wrong. In light of these considerations, I recommend the following auxiliary proposition to replace A2:

> A2": For any act, x, if x is the deliberate killing of a person against her will in a non-self defense and non-doomsday scenario, then x is not morally right.

A2" is close enough to a plausible moral principle for our purposes. When conjoined with A1, it entails the following empirical prediction:

> A3': For any act, x, if x is the deliberate killing of a person against her will in a non-self defense and non-doomsday scenario, then x does not maximize hedonic utility.

7.3.2. <u>AUh predicts unsuccessfully.</u>

With the trouble over the auxiliary premise of example A settled, we can now turn to a more serious problem: AUh is more likely to be disconfirmed by the prediction A3' than to be supported by it. It is a longstanding objection to AUh that it is too indiscriminate in countenancing as morally right the deliberate killing of persons when such killing would maximize hedonic utility. Typically, this objection simply points to possible worlds in which the killing of a person against her will (even in a non-doomsday, non-self defense situation) would maximize the balance of pleasure over pain. However, if it should happen that no such possible world is near to our own, then the naturalistic realist who defends AUh might be able to escape the present worry. After all, it is observations of

the *actual* world that confirm or disconfirm empirical predictions. Unfortunately, there isn't very good reason to think that it never happens in the actual world that killing a person against her will maximizes hedonic utility in non-doomsday scenarios. In fact, an anti-utilitarian who is hell-bent on giving the theory an empirical refutation could arrange to make the prediction come true herself! She could arrange for someone who gets intense joy from killing to kill an isolated and very depressed hermit.[6]

The trouble with example A, then, is that it is a case in which a representative moral theory is likely to be refuted by its empirical prediction, rather than supported by it. Thus, it is not an example that could be used to support the ultimate moral argument.[7]

7.3.3. Example A restated using a different moral theory.

Those who are inclined to retain the auxiliary hypothesis A2" will perhaps doubt that AUh really is our best current moral theory. As it so happens, neither Sturgeon, nor Brink, nor Boyd is a proponent of AUh. As we saw in Chapter 5, Boyd and Sturgeon endorse a view called 'homeostatic consequentialism.' Brink endorses a view that he calls 'objective utilitarianism.' Let us consider, then, how one of these more sophisticated theories of morally right action fares when it is conjoined with A2" to

[6] Or perhaps it would not be so easy for the malevolent anti-utilitarian: some have argued (usually as an objection to AUh) that it is nearly impossible to make reliable utility calculations about actual act-tokens and their alternatives with any kind of accuracy (see, for example, Lenman 2000). If so, the anti-utilitarian would have a very hard time ensuring that the killing she is planning really will be an act that maximizes hedonic utility. Fair enough. But this only raises a further problem for example A: if it really is that difficult to assess the hedonic utility of an act and its alternatives, then we will find ourselves unable to know whether AUh is vindicated by prediction A3'; we would never know whether or not the act of killing that we observe really maximizes hedonic utility In that case, example A cannot be used to show that AUh is a highly instrumentally reliable theory; we would never know whether its predictions are successful or not.

[7] In fairness to Sturgeon, example A works well enough for his immediate purposes in his (1985a). He presents the example merely to show that moral theories have empirical consequences when conjoined with auxiliary hypotheses. His goal in that paper is not to show that our best moral theories are, in fact, instrumentally reliable.

produce an empirical prediction. Because Brink's objective utilitarianism is better developed and less unwieldy than homeostatic consequentialism, I propose to use it as the sample moral theory for my discussion. Unfortunately, since the value component of objective utilitarianism eludes a concise formulation, and since Brink's own formulation of the view is sketchy and incomplete, my formulation of the theory will be very rough. Here it is:[8]

OU: (1) For any act-token, x, x is morally right if and only if x maximizes objective utility.

(2) The objective utility of thing (e.g. an action, a motive, a state of affairs, etc.), x, is a quantity that increases with the degree to which x "expresses" at least one of the following:
  (a) the pursuit of an admissible project by an agent; or
  (b) the realization of an admissible project by an agent; or
  (c) personal and social relationships that exhibit respect for persons.

(3) For any project, x, and moral agent, S, x is admissible if and only if:
  (a) the formation and pursuit of x by S is *reflective*; i.e., S attempts to integrate x into a "coherent life plan" that realizes S's human essence
  (b) x shows respect for other persons; i.e., S's pursuit of x does not cause other persons "significant and avoidable harm" and involves the recognition that other person's well-being matters.

[8] Objective Utilitarianism is introduced in Brink (1989: ch. 8, especially pp. 231-237, although refinements are made throughout the chapter). My formulation of OU centers around the following passage in which Brink lays out what I take to be the core of his axiology: "As components of human welfare, pursuit and realization of admissible projects and personal and social relationships exhibiting respect for persons are intrinsically valuable. Actions, motives, and other things that express these values are themselves intrinsically valuable, while actions, motives, and other things that causally contribute to the realization of these values are extrinsically valuable" (*ibid*. 234). He conjoins this theory of value with an act-utilitarian theory of right-action: "an action is right just in case it contributes to human welfare at least as much as any alternative action available to the agent" (*ibid*. 237); "…rightness is identical with the maximization of welfare…" (*ibid*. 238). The specification of admissible projects can be found in (*ibid*. 232f). Clause 3b is also fleshed out in two additional passages: "But there are moral constraints on valuable projects; in order for the pursuit or realization of a project to be of value, that project must, among other things, respect other people at least in the minimal sense of not causing significant and avoidable harm" (*ibid*. 264); "…[A]n agent's projects cannot be valuable if they fail to show respect for others. Respect for others requires recognition that others matter, and this requires that we recognize their claims to basic well-being and overall welfare in certain ways" (*ibid*. 289).

There is no doubt that much needs to be said in order to elucidate OU. However, I will confine myself to only three comments. First, my formulation of OU takes at least one liberty with Brink's own presentation of the theory: in order to cast OU as a full criterion of morally right action, I have presented conditions 3a and 3b as jointly sufficient for a project to count as admissible. In Brink's presentation, he explicitly allows that there might be additional constraints on projects; but he declines to outline what they might be (1989: 232). Nothing that I say below turns on whether or not there are additional conditions that a project must meet in order to be admissible. Second, it should be noticed that while clauses 2 and 3 of OU spell out what sorts of considerations give positive intrinsic value to a state of affairs, Brink offers no account of what sorts of considerations, if any, confer negative intrinsic value upon a state of affairs. Finally, the axiological component of OU (constituted by clauses 2 and 3) is to some degree circular. This is because the notion of well-being appears in the specification of the conditions of admissibility for projects. Unfortunately, this last feature of OU complicates things with respect to the claim that it is empirically observable (or at any rate, empirically confirmable) whether an act maximizes objective utility. I propose that we simply overlook this complication and give Brink and the ethical naturalists the benefit of the doubt by assuming that whether an act maximizes objective utility is something that, at least in principle, can be empirically confirmed.

I am not interested in criticizing OU as a theory of morally right action. For the purposes of this discussion, I am ready to grant that it is the best available theory of its kind. Instead, my present concern is with the question of whether OU, when conjoined

with auxiliary premises such as A2", yields enough interesting successful predictions to render a realistic reading of the theory superior to any anti-realist readings.

Now, the conjunction of OU and A2" yields the following (putatively) empirical prediction:

> A3": For any act, x, if x is the deliberate killing of a person against her will in a non-self defense and non-doomsday scenario, then x does not maximize objective utility.

Let us use 'example A-II' to refer to the set of propositions whose members are OU, A2", and A3". In assessing the prospects of using A-II as the foundation of an ultimate moral argument, we must begin with the question of whether the prediction captured in A3" is true. Again, because A3" is a universally quantified proposition, no single observation can confirm *it*. What we are interested in is, rather, whether it is the case that all observed acts of killing a person (against their will and in a non-doomsday scenario)[9] have been acts that fail to maximize objective utility.

I suspect that few people have ever, upon witnessing a killing, stopped and tried to discern with any accuracy whether or not no alternative action expressed to a greater degree the pursuit or realization of projects or relationships that exhibit respect for persons. Thus, I doubt that the kinds of predictions yielded by A3" have been directly confirmed by empirical observation. Nevertheless, I am ready to concede that a plausible case can be made that such predictions are likely to be vindicated. Here is how such a case might go: Any act that involves the deliberate killing of a person against their will can plausibly be argued to be an instance of pursuing a project that fails to "show respect for other persons." In particular, it can plausibly be maintained that all such acts cause

[9] From here on, let it be understood that by 'acts of killing a person,' I mean 'acts of killing a person against their will in a non-doomsday scenario.'

“significant and avoidable harm” of the kind that Brink takes to be constitutive of failing to respect other persons (Brink 1989: 264). If so, then all actions that involve killing persons fail to express the pursuit or realization of an admissible project or a relationship that exhibits respect for others. From these considerations, it follows that the objective utility of any such action is, at best, 0. A defender of OU might be able to argue successfully that, in all actual world cases, there will always be some other alternative open to an agent that has greater objective utility. Because of this, we have good reason to think that the prediction that all acts of killing a person fail to maximize objective utility will ultimately be born out.

7.3.4. <u>The approximate truth of OU is not needed to explain its predictive success.</u>

Let us grant, then, that A-II is example of a successful empirical prediction by what is our best theory of morally right action. This is a start towards constructing a viable ultimate moral argument; but it is not enough. If A-II is to ground such an argument, then it must also be the case that the best explanation of our ability to predict A3" using OU requires us to hold that OU is (approximately) true. I think there is reason to doubt that the best explanation really requires this. Let me explain.

Unlike cases of scientific prediction, the relevant moral theory (OU conjoined with A2") offers no explanation of *why* A3" obtains. Those of us who think A3" is true certainly do not think it is true *because* all acts that maximize objective utility are morally right and no act of killing a person is morally right. Instead, we expect that the explanation for why A3" obtains will be of a psychological or sociological sort: e.g., people desire not to be killed; being killed usually conflicts with a person’s projects and

values; etc. (The exact form of the psychological explanation will depend upon how we cash out the notion of harm and well-being in clause 3b of OU). Most importantly, we have every reason to expect that, if there is a fully satisfying explanation of why A3" obtains, it will be one that makes no ineliminable use of moral vocabulary. The upshot here is that we should have a very high degree of confidence that we will be able to satisfactorily explain why A3" is true without having to assume that OU (or the conjunction of OU and A2") is approximately true.

7.4. Example B: Predictions Based on Moral Character.

7.4.1. <u>The independence of the prediction.</u>

The first difficulty with example B concerns the independence of the prediction B4 from the moral theory or principle captured by B1. This worry arises primarily with respect to the role that B1 (or a more general moral theory that implies B1) plays in our arriving at B3, the judgment that Zenobia is morally good. Here is the problem: suppose that our evidence for Zenobia's being morally good is our past observation that she always keeps her promises, even at great personal cost. Notice now that the very same evidence is sufficient to justify the conclusion, simply by way of enumerative induction, that B4 is likely to be true: Zenobia will keep her promise.[10] Given the evidence already in our possession, the knowledge that B1 is true adds nothing to our ability to predict Zenobia's

[10] Below, we will see that there is reason to doubt that such an inductive inference is cogent, at least if we are trying to predict whether Zenobia will keep her promise in a novel situation—i.e., a situation that is importantly different from the past situations in which she was observed to keep her promises. But, as I argue below, these doubts only make things worse for those who would use B1 to make predictions. My present claim, then, proceeds from assumptions about personality and behavior that the ethical naturalist already needs to accept if example B is to be successful in grounding the ultimate moral argument.

behavior.[11] Obviously, then, the approximate truth of B1 is not needed in our best explanation of our ability to successfully predict B4.

Although Brink's own reasons for producing his version of example B do not directly concern the role it might play in a constructing an ultimate moral argument, he does notice the threat of dependence between B3 and B1. To address it, he suggests that our evidence for B3 ought to be independent of Zenobia's promise-keeping behavior (1989: 137). This requirement can be satisfied, perhaps, if the judgment that Zenobia is a morally good person is arrived on the basis of observations that reveal her to have virtuous character traits that are suitably distinct from promise-keeping. For example, we might have judged Zenobia to be good in response to our observation that she is (or that she generally behaves in ways that are) kind, courageous, and/or temperate. For this maneuver to be plausible, it must be the case that we are warranted in inferring that a person is morally good merely from the observation that she is kind, courageous, and temperate, etc. Thus, for the purposes of building an ultimate moral argument, we must see B1 as part of a broader theory of moral character according to which characteristics besides promise-keeping or *fidelity* (e.g., characteristics such as *courage*, *temperance*, *kindness*, etc.) are constitutive of a person's being morally good. But more than this, it also requires such a theory to incorporate a kind of "unity of the virtues thesis." This unity thesis need not be as strong as some traditional versions whereby a person counts as morally good only if she has *all* of the good-making character traits (i.e., virtues). But it

[11] It is true that, since the observation of Zenobia's past promise-keepings have led us to judge her to be morally good, we must already be committed to accepting a moral theory that incorporates something like B1. But this fact does not show that the approximate truth of B1 is needed in the best explanation of our success in predicting Zenobia's behavior. If, instead of accepting B1, we thought that keeping promises was a bad character trait to have, we would still have been able to successfully predict Zenobia's behavior by enumerative induction on the basis of our observation of her past promise keepings. The only difference in this case is that is that, instead of accepting B3, we would have judged Zenobia to be morally bad.

does require, at the very least, that there is a strong nomological connection between the having of some of the virtues and the having of the rest. (In other words, the virtues must be homeostatically clustered). This connection must be strong enough so that the observation that a person is kind, temperate, courageous, etc., warrants an inductive inference to the proposition that that person is also disposed to keep promises, even at great personal cost.

7.4.2. Trouble from social psychology.

On the face of it, it seems plausible enough that a person who is kind, courageous, and temperate is also the sort of person who honors her promises even at great cost to herself, and that this latter disposition will be reflected in her behavior on the occasion that we are trying to predict. Unfortunately, there exists a body of psychological research that casts doubt on the cogency of such an inference. This research suggests that stable character traits are much less predictive of agent's behavior than are the features of the particular situation that the agent finds herself in. Among the classic studies that support this conclusion is one by Alice Isen and Paula Levin (1972). In one experiment they found that subjects were significantly more likely to offer help to a person who had dropped papers on the floor of a shopping mall if, moments before, the subject had found a dime in the coin return of the public phone from which she had just made a call. Of the subjects who did not find a dime after checking the coin slot, only 4% stopped to offer help. By contrast, 87% of those who found a dime in the coin slot offered help. This data supports the conclusion that, in this sort of case at least, facts about the agent's situation (such as whether she found a dime) are a greater predictor of her behavior than

are facts about the agent's moral character. Other classic studies that have been invoked to support this "situationist" view of behavior are Stanley Milgram's (1963) famous obedience study and the Stanford prison simulation (Haney, et al. [1973]). In his attack on virtue ethics, the philosopher John Doris argues that these and other studies establish, among other things, that

> Personality is not often evaluatively integrated. For a given person, the dispositions operative in one situation may have an evaluative status very different from those manifested in another situation; evaluatively inconsistent dispositions may "cohabitate" in a single personality" (Doris 2002: 25).

In other words, virtuous character traits such as *kindness*, *courage*, and *temperance* are not strongly connected to other virtuous character traits such as, e.g., *honesty* or *fidelity*. If this is correct, then we should not expect that the judgment that Zenobia is good—when made on grounds other than her honesty or promise-keeping tendencies—will be of help in predicting how she will behave when her fidelity to her promises is tested. But in that case, it is hard to see how those who want to make use of example B can maintain the requisite independence between theory and auxiliary hypotheses without undermining their ability to make a successful prediction about how Zenobia will act.

But things may be even worse for example B. Doris argues that these studies show more than just that that we cannot accurately predict a person's behaving in accordance with one kind of virtue (e.g. honesty) in a particular situation on the basis of evidence that they possess virtues of a different kind (e.g., courage, charitableness). The studies also show that, in general, we cannot predict how an agent will behave in a novel situation in which a virtue is tested, *even if that agent has always acted in accordance with that virtue in more familiar situations* (Doris *ibid*.; Nisbett and Ross 1991). In other words, upon observing that an agent has always behaved honestly in her personal

dealings (even when there has been temptation to lie), we cannot reliably predict that she will also behave honestly in, say, a business setting. Suppose, then, that we have judged Zenobia to be good precisely because we have observed her always to keep her promises. If the situationist account of behavior is correct, then we will not be able to predict that she will keep her promise the present situation *unless this situation is relevantly similar to others in which we have observed her to keep her promise*. If the situation with respect to which we are trying to predict Zenobia's behavior is one in which she has made a promise to keep a secret—and not one in which she has promised to look after a friend's child, to repay a debt, to be faithful to her spouse, to donate her kidney, etc.—then we should expect our prediction to be reliable only if we have previously observed Zenobia to keep her promises not to tell secrets. However, we cannot reliably predict she will keep this promise if our basis for this prediction is merely our past observations of her having always kept her promises to care for her friend's children.

Although I cannot here defend the situationist account of human behavior from its critics (for example, Epstein and O'Brien 1985), the very presence of a vibrant situationist research program in social psychology should be sufficient at least to raise significant doubts about how useful example B is as a foundation for the ultimate moral argument. If situationism is correct, then example B-type predictions—predictions about how a person will behave based on observations about their moral character—are not likely to prove reliable and accurate. Consequently, example B is not likely to be an example of a successful empirical prediction. Thus, it does not offer promising evidence that our best current moral theories are instrumentally reliable to such a high degree that their reliability is best explained only by supposing them to be approximately true.

7.4.3. A competing non-moral explanation of the predictive success of B1.

For a defense of the situationist theory in social psychology, the best I can do is to refer readers to Ross and Nisbett (1991). Since I have neither the space nor the expertise to mount such a defense myself, it might best that I do not rest the entirety of my case against example B on the success of situationism. Fortunately, example B fails as a foundation for an ultimate moral argument for additional reasons that are independent of situationist considerations.

To see why example B fails, we will need to make several assumptions, all of which are concessions that favor the ethical naturalist in this context. To begin, let us assume that there are character traits such as *fidelity*, and that persons who have these traits really do exhibit an unwavering tendency to behave in accordance with them across all the kinds of situations that human being normally encounter. Let us grant, further, that several of these character traits are homeostatically clustered. And let us suppose that there is at least one clustering of such traits such that all of the traits in that cluster are ones that we commonly think of as virtues. To fill out this assumption with some detail, let us suppose that the properties or character traits *fidelity*, *courage*, *temperance*, *humility*, and *kindness* are homeostatically clustered for human beings: if a person has four of these traits to a high degree, then there is a much greater than average likelihood that they will have the fifth trait to at least a significant degree. Finally, granting that these five properties are natural properties, I will use the non-moral term 'N_1' to denote this natural property cluster.

Above, I suggested that B1 should be viewed as a part of—or else as an implication of—a more general theory of *morally good personhood*. We can now see what such a theory would look like. If the above assumptions are granted, then the following moral theory could be used to underwrite B1:

GP: a person, S, is morally good if and only if S instantiates N_1.

The question before us, as I see it, is whether the best explanation for our ability to successfully predict B4 on the basis of GP and the auxiliary premises B2 and B3 requires us to view GP as approximately true. Can we explain the predictive success of GP just as well if we suppose that GP is false, and not even approximately true? I think we can.

In order to explain how it is that proponents of GP are able to successfully predict B4, we need only to take on a commitment to the existence of the property cluster N_1 and a commitment to the proposition that proponents of GP identify *morally good personhood* with N_1. These commitments do not require us to suppose that this identity claim is true. What explains the ability of GP proponents to predict B4 is simply (i) the fact that they ascribe *morally good personhood* to all and only persons who instantiate N_1 and (ii) the fact that all persons who instantiate N_1 always keep their promises, even at great cost to themselves. I see no reason to think that this explanation is any worse than an explanation that takes GP to be approximately true. Thus, the predictive success of GP does not commit us to realism about *morally good personhood*.

There is a potential objection to this line of argument that I need to address. The objection is that my supposed explanation of GP's success is not really a non-moral explanation, as I claim. The trouble is that, in describing N_1, I made use of the terms 'fidelity,' 'courage,' 'temperance,' 'humility,' and 'kindness.' These terms are often

thought of as part of our moral vocabulary. Perhaps one reason for this thought is that at least part of the linguistic function of these terms is evaluative: under normal circumstances, to call someone courageous is to praise her. Unless there is a satisfying explanation of GP's predictive success that can be stated entirely without recourse to moral vocabulary, it looks like we will find ourselves committed to moral realism after all.[12]

My own view is that, in their typical uses, terms like terms like 'fidelity' and 'courage'—the so-called "thick" moral terms—perform a dual function: first, they pick out a syndrome of behavioral or psychological dispositions; and second, they function either to praise the having of these dispositions, or else, to represent the possession of them as praiseworthy. I believe that in certain contexts, these terms can be used in such a way that their second, evaluative function is suppressed or canceled. When one of these terms is used in such a way, I contend, it no longer functions as part of moral vocabulary, but rather, as part of the vocabulary of psychology or some other social science (cf. Blackburn 1998: 101-104; Gibbard 1990: 112-115; Hare 1963: 24f). As evidence that thick moral terms can be used in non-evaluative ways, consider the fact that we sometimes debate whether the character traits they denote really are virtues, i.e., whether or not they really are praiseworthy. Now, in order to render GP plausible, I chose character traits whose status as virtues are unlikely to be disputed by contemporary secular ethicists (although I would not be surprised if some doubt whether *humility* is a virtue). Consider, however, the following character traits, all of which have at one time or another been suggested to be virtues by speakers in our intellectual tradition:

[12] Sturgeon raises a similar objection against Harman's supposed non-moral explanation of the judgment that setting the cat on fire is wrong (Sturgeon 2006a: 251).

cleanliness, *chastity*, *frugality*, *humor*, *industriousness*, *loyalty*, *manliness*, *modesty*, *patience*, *patriotism*, *pride*, *religious faithfulness*, *selfishness*, and *tolerance*. It seems to me that there is nothing at all odd about asking of any item on this list whether it is a virtue. If I am right, then there is no reason why the terms used to refer to these character traits cannot be construed as part of non-moral vocabulary, at least if we decide that the trait it picks out is not, upon reflection, a virtue.[13] And if we allow this, then there seems to be no reason why an ethical nihilist (who holds that no character trait has the property of *being a virtue*) cannot utilize these terms as part of a lay psychological vocabulary that functions to pick out interesting behavioral and psychological dispositions. It seems to me, then, that there should be no trouble with construing the five terms used to describe N_1 as non-moral or non-evaluative, as I intended them.

For those who remain unconvinced, however, I recommend a different strategy: when describing N_1, we should simply replace the thick moral terms with uncontroversially non-moral descriptions that pick out the very same character traits. For instance, in place of the suggestion that one component of N_1 is the property *courage*, we may instead pick out the relevant component property using the term 'being disposed to remain calm and be steadfast in the face of great danger.' This latter term, I take it, is uncontroversially non-moral. In addition, it picks out the same character trait that most

[13] Note that I am not claiming—nor do I need to claim—that all putatively thick terms can be purged of their evaluative import. I have doubts about whether one can use a racial epithet or ethnic slur non-ironically without being understood by others to be denigrating people of the relevant race or ethnicity. The evaluative function of such terms is not cancelable in our language, as our language currently is. But even if this is true for epithets and slurs, I do not think the same must be true for terms denoting character traits. (On the other hand, it is worth observing that epithets sometimes evolve so as to lose their negative connotations. This is arguably the case with the term 'queer,' which, despite seeing no significant change in the extension that it picks out, no longer implies a negative evaluation of those to whom it is applied, at least in many common contexts. I hesitate to push this point too far, since it might be argued that these uses of 'queer' are merely ironic. Still, it does seem to me that the term certainly is on a trajectory where it may eventually have an evaluatively neutral use that is perfectly earnest.)

uses of 'courageous' pick out.[14] At any rate, I believe it comes close enough to give us confidence that some nearby non-moral property term will do the job.[15] Utilizing this strategy, I believe it is possible to offer an explanation of the predictive success of GP using purely non-moral vocabulary.

---

[14] It might be objected that thick moral terms denote more than just a syndrome of behaviors. What makes the term apply to those behaviors is that those behaviors are either *appropriate* or *inappropriate*, where 'appropriate' and 'inappropriate' are moral terms, each denoting an irreducible moral property. For example it might be said that 'cowardice' does not denote (something like) the property of *feeling fear in situations that do not cause the average person to feel fear*; instead, it denotes something like the property of *taking fear in things that are inappropriate objects of fear*. If this is so, then we cannot replace 'cowardice' with a thoroughly non-moral, non-normative term. (And the problem, of course, is that it would be difficult to resist the conclusion that something similar holds for virtue terms, like 'courage' and 'fidelity'.) Whatever the merits of this kind of argument, it does little good for the naturalist's cause. The problem is that what explains and predicts the coward's behavior (e.g., her fleeing from the sight of a spider) is the fact that she is disposed to take fear in objects of a certain sort (e.g., spiders, or animals, or things that bite, etc.) and not the fact that these objects are inappropriate objects of fear. That is to say, if *cowardice* is an irreducibly normative or evaluative property, then *cowardice* is never what explains or predicts a person's behavior. Similarly for other thick moral terms that are given this sort of analysis.

[15] I should acknowledge that some philosophers have expressed doubts about the prospects for such a maneuver to succeed. For example, John McDowell writes, "It does not follow from the satisfaction of [the requirement that evaluative classifications are supervenient on non-evaluative classifications] that the set of items to which a supervening term is correctly applied need constitute a kind recognizable as such at the level supervened upon. In fact supervenience leaves open this possibility…: however long a list we give of items to which a supervening term applies, described in terms of the level supervened upon, there may be no way, expressible at the level supervened upon, of grouping just such items together" (McDowell (1981/1998: 202). To avoid confusion, I should make it clear that my proposal does not require that the extension (or, perhaps better: intension) picked out by the relevant non-moral term forms a natural kind from a non-moral point of view. All it requires is that we can pick out that intension using non-moral vocabulary. To do this, it is not required that the relevant intension corresponds to a natural kind from the non-moral point of view. After all, most agree that there is no natural kind corresponding to 'grue'; but for all that, it is not controversial that there is an intension that 'grue' picks out.

The more serious charge against my proposal, then, is that the natural property that 'courage' picks out has such a wildly heterogenous extension that it is simply beyond human capacity to describe this extension from anything other than a moral perspective. I think it is plausible that this kind of situation arises with respect to the vocabularies of some sciences or disciplines that describe a certain class of facts and the vocabularies of those disciplines that describe the facts upon which the former facts supervene. I would not be surprised, for instance, if there is no humanly possible way to pick out the intension the psychological predicate 'pain' using only the vocabulary of fundamental physics. But that is not the situation that we are in with respect to moral vocabulary. I am not suggesting that we pick out the natural property denoted by 'courage' by limiting ourselves to the vocabulary of fundamental physics; I am suggesting that we do so using (primarily, but not exclusively) the resources of personality psychology. Of course, it could turn out that when we attempt a fully adequate non-moral description of the kind of person to whom 'courageous' is applied, we will find that the vocabulary of psychology and the rest of the sciences simply aren't up to the task. It seems to me, however, that the more natural position here is the optimistic one.

7.5. Example C: Predictions Grounded in a Causal-Moral Generalization.

7.5.1. Does justice really cause social stability?

Sturgeon himself locates the major weakness of example C. The plausibility of C rests on its auxiliary premise, C2. According to C2, the justice of a society implies (because it causes) social stability. Sturgeon acknowledges, however, that "there is…a tradition that attacks this latter claim as a pious fiction" (1992: 106). While I would not quite call C2 a pious fiction, I do think there is reason to doubt it. As Brian Leiter writes, "it seems that justice provokes opposition as often as it produces allegiance: many people have little interest in just arrangements, and so resist them at every step" (Leiter 2001: 95). Indeed, if, as I believe, a just society in the United States would more closely conform to RPJ than it does to its present socio-political arrangement, then it would be nothing short of naïve to suppose that movements towards justice would strengthen—rather than weaken—social stability, at least in the short term. In order to satisfy Rawls's principle 2a (the so-called "difference principle"), there would surely need to be far greater taxation on the wealthiest individuals for the purposes of redistribution. Unfortunately, there is a sizable constituency in the United States (which includes, not surprisingly, many of its wealthiest citizens) who view heavy taxation for such purposes as an affront to liberty and as deeply unjust. These citizens would surely offer vocal—and very possibly destabilizing—resistance to any proposal to enact policies that would push the U. S. towards greater conformity with RPJ. If this is correct, and if Rawls's theory is our best theory of justice, then we cannot accurately predict that a society will be stable upon our observing that it satisfies RPJ.

7.5.2. Is C2 a non-negotiable condition of adequacy for theories of justice?

Of course, if a society that is observed to satisfy RPJ is found to be unstable, then, rather than reject C2, we might reject Rawls's theory of justice instead. Indeed, Sturgeon suggests that we could use C2 as a means for selecting between alternative theories of justice. By this methodology, if there is some plausible rival theory of justice, T, and if societies conforming to T exhibit stability (whereas societies conforming to RPJ do not), then we have reason to think that T, rather than Rawls's theory, is true (cf. Sturgeon 1991: 29). Rawls himself appears to accept something like this as part of his own methodology:

> "It is…a consideration against a conception of justice that, in view of the laws of moral psychology, men would not acquire a desire to act upon it even when the institutions of their society satisfied it. For in this case there would be difficulty in securing the stability of social cooperation" (Rawls 1971/1999: 119).

It needs to be recognized, however, that both Rawls and Sturgeon suggest only that the tendency to stabilize is *just one* consideration among others for selecting a theory of justice. Neither philosopher denies that there might be competing reasons with enough strength to justify our acceptance of a theory of justice that would not have a stabilizing effect when it is satisfied. In other words, under certain circumstances, social instability is an acceptable theoretical (and practical) cost of a theory of justice. But if we accept this possibility, then we have rejected C2. Consequently, it is not clear, after all, that Rawls (and, for that matter, Sturgeon) really accept C2; both philosophers allow the possibility that our best theory of justice fails to exert a stabilizing influence on society.

But suppose it is argued that C2 is a non-negotiable part of our criterion of theory selection when it comes to theories of justice. By this way of thinking, we must reject any principles of justice that fail to make (or keep) a society stable when that society

satisfies them. If we take this stance, then it all but guaranteed that our best theory of justice will make it possible to predict that certain societies are stable.[16] But again, this approach compromises the independence of the moral theory from the relevant auxiliary hypotheses: C2 would become an important premise in the argument that leads to the construction and acceptance of our best theory of justice, whatever that turns out to be. Consequently, any prediction concerning the stability of a society on the grounds that it is (un)just would not be a *novel* prediction. Because of this, we can again expect to have serious doubts about whether a realistic construal of the best theory of justice is really needed to explain its success in "predicting" the stability of societies that satisfy its conditions.

More importantly however, I think there is no good reason to hold C2 to be a non-negotiable condition of theory selection. I certainly agree, of course, that we should hope that the best theory of justice is such that the satisfaction of its principles has a stabilizing effect on societies. At any rate, we should hope that our best principles of justice do not have a destabilizing effect on societies. Since most of us very strongly want societies to be *both* stable and just, it would be a shame if we had to choose between these desiderata. Be that as it may, I see no reason why we should think that *justice*-making properties must be stabilizing. Consider, first, a libertarian capitalist theory of justice of the sort advanced by Robert Nozick in his *Anarchy, State and Utopia* (1974). The human tendency to experience envy being what it is, it may be that libertarian capitalist principles of justice, which permit radical economic inequality, simply cannot be satisfied by a society without resulting in significant social instability. I suspect that this could be

[16] Or, at any rate, it is all but guaranteed that those societies will exhibit more stability than societies satisfying competing principles of justice.

true even if the economy of such a society succeeded in meeting the basic needs of the worst off. If this is so, then this strikes me as a reason not to want to see such principles of justice realized. But this does not strike me as a reason to think that those principles must be *false*. I think something similar holds for the sorts of egalitarian principles of justice that I favor. Human nature being what it is, it could turn out that egalitarian principles of justice would never be accepted by large segments of the most powerful groups in any given society. If this were true, then such principles would make for a less stable arrangement than some other, less egalitarian principles of justice. While such a fact might lead me to hesitate as to the matter of whether, all things considered, I would like to see egalitarian principles of justice realized, I am less inclined to worry that the instability would indicate the incorrectness of those principles. It may simply be an unfortunate fact about human nature that the most just arrangement can never garner as wide an allegiance as some less just arrangement. For this reason, it seems to me that C2 should not be held as a non-negotiable condition of adequacy for theories of justice.

7.5.3. Justice is not what explains stability.

Finally, it seems to me that, to the extent that there is a causal or explanatory relationship between *justice* and *stability*, what explains the stability of a society is not the fact that the society satisfies principles of justice that are stance-independently true; rather, what explains the stability is the fact that the society satisfies principles of justice that are widely *accepted* by its members. As long as a large enough segment of society can and does internalize and subscribe to its principles of justice—whatever they happen to be—it seems plausible to suppose that the society will be to that extent stable. If this is correct,

then, unless a plausible case can be made for the claim that only the *true* principles of justice can be internalized by a vast supermajority of denizens living within the boundaries of a state—and I see no reason to think such a case will succeed—we should expect that even unjust states can exhibit a good deal of stability.[17] But in that case, when a theory of justice, conjoined with C2, successfully predicts that a society is stable, this success can be satisfactorily explained without supposing that the theory is approximately true. Thus, example C fails to show that the approximate truth of our best theory of justice is needed in order to explain the instrumental reliability of that theory.

7.6. Moral Explanations and Interesting Generalizations.

There is a final point that I need to address, although it is not entirely obvious to me where it fits in the current dialectic. Brink and Sturgeon suggest that there are some moral explanations of non-moral facts that cannot be replaced by wholly non-moral explanations that cite instead only the non-moral supervenience base facts that realize those putative moral facts. This sort of situation might occur when there are a number of distinct non-moral properties that can realize the same relevant moral property in different circumstances. It may be that the instantiation of the moral property would have

[17] Against this, Brink suggests that sometimes "there will be cases where the causal efficacy and explanatory power of moral facts precede their recognition" (1989: 189; cf. Railton 1986: 192). Unfortunately, although Brink offers an illustration showing how this *might* happen, he offers nothing in the way of evidence to suggest that it ever *actually does* happen. Moreover, the illustration that he offers (which involves a person who comes to unreflectively resent his social position despite his accepting an inegalitarian ideology according to which his own inegalitarian society counts as just) does not obviously preclude a satisfactory non-moral explanation. Again, for example, it may simply be a fact about human nature that we resent it when some have more power and goods than we have. (And this feature of human nature may be so recalcitrant that it operates even when we consciously accept inegalitarian principles of justice.) If this was a fact of human nature, then it may well be true that the inequality of social arrangements causally contributes to resentment among those of a lower socio-economic status. But we should expect this resentment to occur regardless of whether the true principles of justice are, in fact, egalitarian ones. In fact, we should expect this resentment regardless of whether any principles of justice are true at all.

the same causal consequence no matter which of its potential non-moral realizer properties has realized it. In that case, it could be argued that a better explanation of the effect in question would cite the higher-level, moral property (or the fact of its instantiation), rather than the lower-level, non-moral realizer property (Brink 1989: 193-197; Sturgeon 1998: 201; 2006: 251f).

One reason that I am unsure how this fits into the present dialectic is that it is not clear how this consideration, if true, could be used to support an ultimate moral argument. I suspect that it could support such an argument if it should turn out that theories that recognize the higher-level moral property make better, more accurate predictions than any theory that does not recognize it. If so, then perhaps we will need to cite the approximate truth of those theories in order to explain their predictive success. However Brink and Sturgeon's claim relates to the prospects for building an ultimate argument, I want to address their suggestion that, because explanations citing only non-moral realizer properties might fail to capture "interesting generalizations"—generalizations that a better explanation would illuminate—our best explanations of non-moral facts might require us to make reference to moral properties.

I do not deny that Brink and Sturgeon are right when they suggest that if no non-moral explanation can capture the right generalizations or support the right counterfactuals, then moral facts would be needed in our best explanation of some state of affairs or event. I do not believe, however, that they have shown that it is likely that there really are regularities or generalizations or other phenomena for which a moral explanation is better than all competing non-moral explanations. To begin with, consider explanations couched in terms of an agent's moral character. In §6.2.2 I noted that that

Sturgeon maintains that Hitler's genocidal actions are explained by the fact that he was morally depraved. In that same section, however, I pointed out that any explanation that did not go on to cite Hitler's non-moral, depravity-making properties would be a poor explanation. This is so because it isn't true that Hitler would have ordered genocide if his depravity had been constituted by the non-moral properties of *being dishonest* or of *being a pedophile* rather than the non-moral properties of *being homicidal* or *being sadistic*. But once we have an explanation couched in terms of one (or both) of these latter properties, that explanation is not improved by adding the further claim that Hitler was morally depraved. And indeed, such an addition would make for an inferior explanation insofar as it renders the explanation less parsimonious than it would otherwise be. Consequently, it seems that our best explanation of Hitler's actions does not require that we cite his *depravity*.

Now the Hitler example is just one putative moral explanation. So perhaps some better example can be found. I contend, however, that the failure of the Hitler example places the burden of evidence on those who would claim that some moral explanations capture interesting generalizations that cannot be captured by wholly non-moral explanations. Let us examine, then, the sorts of examples that the synthetic ethical naturalists offer to support the claim that some moral explanations cannot be replaced without explanatory loss.

Between the writings of Brink, Boyd, and Sturgeon, I am aware of only one example offered to illustrate a case in which an explanation of a non-moral fact that cites

only non-moral realizer properties is explanatorily inferior to a moral explanation.[18] The example is Brink's:

> …[R]acial oppression in South Africa consists in various particular social, economic, and legal restrictions present in South African society. Now, it seems better to cite racial oppression as a cause of political instability and social protest in South Africa than the particular social, economic, and political restrictions, precisely because there would still have been racial oppression and instability and protest under somewhat different social, economic, and legal restrictions, and the only thing this large set of alternate possible social, economic, and legal bases of oppression have in common is that they realize racial oppression (it is very unlikely that there is a natural – nonmoral – social category that corresponds to this set) (1989: 195).

This example is unpersuasive. In the first place, it is, again, doubtful whether the oppressiveness or the injustice of South Africa's policies during the relevant period of time can explain the protest and instability independently of facts about what South African blacks (and sympathetic whites) *thought* about the system. In other words, what explains the protest and instability is the fact that the socio-political arrangements violated principles of justice that blacks and sympathetic whites *accepted* or *believed*. If so, then the protest and instability is to be expected regardless of whether those beliefs were true. But then, the actual oppression or injustice is not what does the explaining. As support for the claim that what explains the instability are beliefs about what is unjust, rather than *injustice* itself, notice that most of the world's population throughout human history has lived under political arrangements that are, by our own lights, seriously unjust or oppressive. Liberal democracy, after all, is a fairly recent invention. Most humans living under the jurisdiction of a state have lived under some form of aristocracy, monarchy, or oligarchy. I take it that according to our own best theories of justice, all of

[18] To avoid misunderstanding: when I say that the relevant explanation cites only non-moral realizer properties, I do not mean that it cites no other non-moral properties (e.g., non-moral properties that no one thinks realizes any moral property). I mean only that the explanation cites no moral properties in addition to the non-moral realizer properties.

these forms of government are unjust. Moreover, most states that have existed throughout human history have been deeply sexist. The oppression of women was often explicitly written into law, and when it wasn't, it was condoned anyway. Despite the injustice of these arrangements, many of these states managed long eras of internal stability. It seems to me that if injustice or oppression—rather than merely the belief that one's society is unjust or oppressive—is what explains political instability, this historical fact would be more surprising than it is.

Secondly, I do not concede Brink's claim that there is no non-moral property realized by the various possible social, economic, and legal arrangements that are said to be potential realizers of *racial oppression*. It is true that at present we may not have a convenient non-moral term for such a property; but I see no reason to think that such a term cannot be concocted. As a first approximation, consider the following non-moral term: 'being a socio-political system that allocates basic rights and privileges differentially on the basis of race.' For convenience, let's abbreviate this term with 'N_2.' Like *racial oppression*, the property picked out by 'N_2' (*viz*., N_2) can be realized by a multitude of different social, economic, and legal arrangements. Indeed, it seems to me that any particular socio-political arrangement that realizes *racial oppression* also realizes N_2. A more difficult question, however, is whether every socio-political arrangement that realizes N_2 also realizes *racial oppression*. I think there is reason to doubt that they do. The trouble is that political systems that employ affirmative action programs to redress past racial injustices can arguably be said to allocate rights and privileges differentially on the basis of race. If this is right, then those political systems will realize N_2. Nevertheless, many of us do not view affirmative action programs to be unjust. Thus, N_2

will not be perfectly coextensive with *racial injustice* as this latter property is characterized by our best current moral theories.

I do not think this feature of N_2 constitutes a serious problem with my objection to Brink's argument. In the first place, if we were really convinced that only a natural property that was perfectly coextensive (or co-intensive) with *racial injustice* can capture the right causal generalizations, then we could simply find a more complex non-moral term to replace 'N_2', one that picks out a natural property that better mirrors the extension (or intension) of 'racial injustice' than N_2 does. I see no reason to be pessimistic that a term of this sort can be found. But secondly, I am not convinced that only a natural property that is perfectly coextensive with *racial oppression* can do the explanatory work that we need of it. In fact, a case can be made that reference to N_2 makes for a superior explanation of the instability of certain societies. There is anecdotal evidence that suggests that government programs allowing for the differential allocation of rights and privileges on the basis of race may cause instability *even when they are employed with the goal of redressing past racial injustices*. The anecdotal evidence of which I speak is the resentment expressed by whites in the United States who complain of "reverse racism" in response to affirmative action programs that favor non-white minorities. Granted, these resentments are not expressed so loudly or intensely that we should say that they constitute or herald full-blown social instability or social protest. But it seems to me that these resentments are no different in kind from the sort of resentments that ultimately lead to mass protest movements. If I am right, then it may well be that reference to N_2 makes for an even better explanation of social instability and social protest than does reference to *racial oppression*.

I conclude, then, that naturalist moral realists have yet to provide a compelling example of a moral explanation of a non-moral fact that cannot be replaced without explanatory loss by a non-moral explanation. We may conclude, further, that to the extent that our best theory of justice successfully predicts which societies will be unstable, we can satisfactorily explain this predictive success without supposing that that theory is true. Nor is there any reason to think that a realist explanation would be better.

7.7. Conclusion.

The prospects for constructing a compelling ultimate moral argument in defense of moral realism are dim. An argument of this sort consists of two premises. First, it is claimed that our best current moral theories are highly instrumentally reliable—where this reliability is a measure of their success at making accurate empirical predictions (with special weight given to those predictions that are novel). Second, it is claimed that the best explanation of this success requires that we suppose those moral theories to be approximately stance-independently true.

We have considered three examples of predictions derived from moral theories. None of these have furnished us with a compelling example of a (i) currently best moral theory that (ii) yields successful empirical predictions, where (iii) the best explanation of this success requires a realistic interpretation of that moral theory. In the absence of convincing examples of predictions that have these characteristics, there can be no compelling ultimate moral argument. While my discussion has obviously not considered all of the possible examples of moral prediction that might be offered, I believe that the (mis)fortunes of A, B, and C give us reason to be pessimistic that any others will do significantly better.

Let me conclude by tying Chapters 6 and 7 together. In Chapter 6, I argued that the most plausible *a posteriori* explanations of our accepting the moral theories that we accept (e.g., the evolutionary story) do not require us to suppose that those moral theories are approximately true, and thus, they do not require us to suppose that there are any stance-independent moral facts. In addition, in the present chapter I have argued that moral facts are not needed in order to explain anything else, such as the apparent predictive successes of our best moral theories or the occurrence of historical events, such protests and revolts. Because metaphysical naturalists are committed to a methodological principle that directs us to accept an ontological commitment only to those entities and kinds that are needed in our best *a posteriori* explanations of observable phenomena, metaphysical naturalists must reject moral realism; and because moral realism is a commitment of SEN, SEN must be rejected as well.

APPENDIX

A DEFENSE OF MORAL TWIN EARTH FROM MISCELLANEOUS OBJECTIONS

In Chapters 2, 3, and 4, I defended the Moral Twin Earth argument against SEN from several objections. In this appendix, I take up several additional responses to the MTE argument that have been made (or that might be made) on behalf of SEN.

A.1. Adopting Partial Non-Cognitivism.

Some have suggested that the appearance of moral disagreement between Earthlings and Twin Earthlings can be satisfactorily accounted for if, allowing that CSN or some other cognitivist semantics accounts for the content of 'morally right' is correct, we adopt a non-cognitivist semantics for the non-moral, "all things considered" 'ought.' Something like this strategy is hinted at by David Copp (2000: 120-124); but it been expressed more explicitly, and developed in greater depth, by David Merli (2002).

Merli distinguishes an "all-in" use of 'ought' from the term's moral, prudential, and aesthetic uses. He sometimes refers to this kind of ought as "the last ought before action." His idea is that, even when it is settled by our moral theory that ϕ is morally obligatory (and thus, that we *morally* ought to ϕ), there remains a further question about whether one ought to ϕ. For instance, there may be cases in which ϕ-ing exacts such a large a cost on an agent that it simply "makes more sense" for the agent to abide by what

he prudentially ought to do, rather than by what he morally ought to do. With this distinction in hand, Merli writes,

> There is, I think, another way of thinking about the last ought before action. This combines realism about moral discourse with expressivism about all-in endorsement. According to this view, moral rightness is a matter of natural fact, but an answer to the question of what to do...is not a factual judgment but an endorsement of one course of action or one set of reasons for action. When I get behind doing the [morally] right thing, I'm expressing my acceptance of certain norms, or urging others to act accordingly, or something along these lines (2002: 236).

This combination of views allows the defender of SEN to make sense of the disagreement between Earthlings and Twin Earthlings without having to deny that 'right' and t-'right' express different natural properties. When an Earthling and a Twin Earthling are debating whether or not to perform an organ harvest, the locus of their disagreement is not about whether the act is morally right; instead, their disagreement is a non-moral disagreement about whether to perform it. Each party is prescribing (or expressing their endorsement of) the action or its omission. Since their disagreement takes place with respect to all-in ought judgments, rather than moral ought judgments, we may still maintain that 'morally right' and t-'morally right' express different natural properties, as is entailed by CSN. Importantly, however, because they have a disagreement in attitude (at the level of all-in ought judgments), we do not need to view the parties as having a merely verbal disagreement. In this manner, it might be argued that the problem of chauvinistic conceptual relativism is avoided. Because of this, the pressure to view Earthlings and Twin Earthlings as expressing the same content with their respective uses of 'morally right' is greatly diminished.

There are two problems with the "partial non-cognitivist" strategy. First, Merli's proposal seems to require that the disagreement between Earthlings and Twin Earthlings,

though substantive, isn't really a *moral* disagreement. However, it is easy enough to imagine a case of disagreement where this result is incorrect. We may easily suppose that the Earthlings and Twin Earthlings find themselves in a situation in which they agree that prudential, aesthetic, legal, and etiquettical considerations are all of negligible relevance to the decision of whether or not to perform, say, an organ harvest; in such a situation, the question of whether to perform the act will hinge entirely on the question what there is most moral reason to do. Unfortunately, as before, it is hard to make sense of the Earthlings' and Twin Earthlings' disagreement about what there is most moral reason to do if we accept CSN; for, if CSN were true, each party's judgment about what there is most moral reason to do will be incommensurable with the judgments of the other party.[1]

The second problem with Merli's proposal is that concedes too much to the moral non-cognitivist opponents of moral realism.[2] Moral realists have long argued against moral non-cognitivism on the grounds that the latter view: (a) requires us to view the declarative surface grammar of moral utterances as misleading, (b) cannot make good sense of moral sentences embedded in conditional statements, (c) cannot make good sense of the apparent logical validity of arguments involving moral predicates, and (d) cannot make good sense of our practice of predicating *truth* of some moral sentences (Brink 1989: 25; 87; 1999: 197ff; Shafer-Landau 2003: 23f). Unfortunately, if naturalistic moral realists adopt non-cognitivism about the all-in ought, they would find

[1] One might be tempted to respond to this objection by proposing that, in this circumstance, their disagreement is over which party's moral norms to all-in accept. Whatever the merits of this kind of maneuver, it seems to me to concede a too much to certain non-cognitivist treatments of moral judgments. In particular, the view that emerges from this way of viewing the parties' disagreement is strikingly similar to the norm-expressivist account of moral judgment advanced by Alan Gibbard (1990).
[2] By '*moral* non-cognitivist' I mean someone who advances a non-cognitivist account to moral judgments, as opposed merely to some other sort of practical judgment (such as all-in ought judgments, or prudential ought judgments, etc.).

themselves obliged to answer these very same objections—only now with respect to all-in ought judgments, rather than moral judgments. To see that this is so, consider the following argument:

1. We (all-in) ought to perform the organ harvest only if we have sterile instruments.
2. It is not the case that we have sterile instruments.
3. Therefore, it is not the case that we (all-in) ought to perform the organ harvest.

This argument is perfectly intelligible. The premises and the conclusion each have a declarative form. In the first premise, an all-in 'ought' appears embedded in the antecedent of a conditional statement. Moreover, the argument appears to be logically valid in the form of *modus tollens*. And finally, it is easy enough to imagine situations in which we will want to say that all of the premises and the conclusion are true. Thus, it should be clear that all of the same phenomena that are thought to raise trouble for non-cognitivism about moral judgments arise with respect to all-in ought judgments as well. This fact poses a dilemma for the proponent of SEN who would avail himself of the partial non-cognitivist solution to Moral Twin Earth: On the one hand, if he cannot answer the standard objections to non-cognitivism, then his solution fails for all the same reasons that moral non-cognitivism fails. On the other hand, if he succeeds in answering the standard objections to non-cognitivism, then he has vastly improved the fortunes of his metaethical rivals, the moral non-cognitivists. Indeed, once it is shown that non-cognitivism about all-in ought judgments is a viable position, it is hard to see why we shouldn't adopt non-cognitivism about moral judgments as well. Given the advantages that moral non-cognitivism has over moral realism with respect to ontological parsimony and an explanation of the intimate (even if contingent) connection between moral

judgment and motivation,[3] the concessions that Merli's proposal requires may well prove fatal for moral realism. At a minimum, however, these concessions deprive moral realism of a great many advantages over moral non-cognitivism that are often claimed on its behalf.

A.2. Brink's Counterfactual Causal Regulation Account of Reference

Before offering his own preferred moral semantics (BMS, discussed in Chapter 4), Brink proposes a revision to Boyd's causal regulation theory of reference. This revision promises to render the application of Boyd's theory of reference to moral vocabulary less vulnerable to MTE-type counterexamples. He suggests that Boyd's account be revised so that we understand causal regulation in the following way: "[A] natural property N causally regulates a speaker's use of moral term 'M' just in case his use of 'M' would be dependent on his belief that something is N, were his beliefs in dialectical equilibrium" (2001: 169).

Because Brink does not ultimately endorse the resulting semantics as answer to MTE, I will not discuss it in depth. I do, however, want to make brief note of two objections to it. First, it can easily be shown that a moral semantics that incorporates this

[3] Naturally, those who accept a strong link between moral judgment and motivation—such as the link expressed by MJI (see §1.5.5)—will find a non-cognitivist treatment of moral judgments more congenial that cognitivist-realist treatment. However, realists such as Brink and Sturgeon deny MJI. Still, there is a weaker kind of internalism that is to my mind less vulnerable, less easy to reject than MJI is. I have in mind a form of internalism that asserts not a necessary link between an individual's moral judgments and her motivation, but rather a necessary link between the having of a moral vocabulary by a community, and the motivational tendencies of its members. Roughly, this form of internalism holds that it is not possible that there be a community of speakers that have a moral vocabulary (and who make moral judgments) where the large majority of these speakers are not regularly motivated to act in accordance with those judgments. We may call this claim *global internalism*. (Something like this form of internalism is suggested by James Lenman [1999: 445f]. He denies the possibility of a form of "global" amoralism, where global amoralism is essentially the contradictory of my global internalism.) The truth of even this weaker form of internalism seems hard to explain given the naturalist moral realist's understanding of moral judgment. Thus, if true, global internalism arguably favors a non-cognitivist construal of moral discourse.

understanding of causal regulation leads to a view of moral facts that is incompatible with moral realism. In particular, it leaves no room for the possibility that our ideal moral theory is false. Here is why: when Brink's counterfactual account of causal regulation is conjoined with CSN, facts about what moral theory we would accept under ideal epistemic conditions themselves fix the referents of our moral terms. This feature of the present proposal violates realism's stance-independence requirement.[4]

A second problem for Brink's proposal is that it is hard to see how the revised understanding of causal regulation is supposed to shield CSN from MTE-type counterexamples. H&T's stipulations are consistent with the assumption that both Earthlings and Twin Earthlings would accept different moral theories even when their respective beliefs are in dialectical equilibrium.[5] Unless this stipulation renders the MTE example incoherent, and I see no reason to think that it does, then even the revised version of CSN wrongly entails that 'right' and t-'right' express different content.

---

[4] Brink appears to acknowledge this in (2001: 175f, especially note 34). In fact, he seems to suggest that genuine moral realism is incompatible with—or at any rate, fits uncomfortably with—Boyd's causal theory of reference as Boyd himself formulates it. (I think this is incorrect; but if I am wrong, then so much the worse for moral realists who would avail themselves of Boyd's semantics.)

[5] While Brink contends that there is no *a priori* reason to think that a common moral theory would not emerge for all rational creatures whose beliefs are in dialectical equilibrium (2001: 170), the proposed response to MTE that we are presently examining requires more than just the *possibility* that there would be convergence in moral belief; it requires that such convergence is *necessary*. The reason why is that, if it is so much as possible that Earthlings and Twin Earthlings fail to converge in the moral theories they accept in dialectical equilibrium, then an MTE-type counterexample can be concocted to refute the revised Boydian semantics. Thus, to sustain the case that the revision of Boyd's semantics avoids MTE, Brink needs to make the much stronger claim that we have good reason to believe that there *could not* be a divergence in moral theory under conditions of dialectical equilibrium. With respect to this strong claim, it seems to me that skepticism is perfectly warranted as a default position.

A.3. Boyd's Achievement Explanation Condition.

A.3.1. The achievement explanation condition and practical success.

On Boyd's most recent formulation of his theory of reference, the mere fact that a term, t, is causally regulated by the properties of a given phenomenon, p, is not sufficient to establish that that t refers to p. What is required, in addition, is that "the epistemic access which uses of t affords [sic] speakers to the real properties of p must (help to) explain the theoretical and/or practical successes achieved in the domains of inquiry or of practice to which t-talk is central" (Boyd 2003a: 515; cf. 538). Boyd calls this the *achievement explanation condition* for reference.[6] Given the addition of this condition to his semantics, it follows that Moral Twin Earth is a problem for SEN only if we can coherently add a further stipulation to the Twin Earth story: where it is stipulated that Twin Earthling uses of t-'right' are causally regulated by T^d properties, we must be able to add the further stipulation that the fact of this causal regulation helps to explain the successes Twin Earthlings achieve in employing moral terms like 'right.'[7] (Similarly, we must be able to stipulate coherently that the causal regulation of 'right' by T^c properties on Earth helps to explain Earthling successes achieved through the use of moral discourse.)[8]

To evaluate whether Boyd's achievement explanation condition poses a threat to Moral Twin Earth, we would need to know what would count as a theoretical or practical

[6] In other contexts, Boyd calls it the *accommodation condition*.

[7] For Boyd, a term's affording us epistemic access to a phenomenon, p, consists in the fact that p causally regulates the use of t. (In the later formulation of his theory of reference, the causal regulation condition is dubbed *the epistemic access condition*, see Boyd 2003a: 538.)

[8] In adding the achievement explanation condition to his semantics, Boyd cannot be accused of making an *ad hoc* modification for dealing with Moral Twin Earth. He adds the achievement explanation condition with the goal of addressing problems of referential indeterminacy that have long posed a challenge to causal theories of reference. His actual motivation for this modification, as far as I am aware, has nothing to do with worries about MTE-type cases.

success of moral discourse. In the natural sciences, it does not seem all that difficult to identify such success. Using scientific theories, we are able to make impressive predictions about observable phenomena. Especially important for the successfulness of a theory are those predictions that are "novel," i.e., those predictions that were not utilized in the construction of that theory.[9] To give an example relevant to the concern at hand, we may plausibly suppose that our scientists' practical achievement of detonating a nuclear bomb (and their ability to predict the conditions under which this achievement is possible) is to be at least partially explained by the fact that the discourse of our best physical theories, which include terms like 'proton,' 'neutron,' and 'electron,' is causally regulated by protons, neutrons, electrons, and their properties. When we turn our attention to the supposed theoretical and practical achievements of moral discourse, however, it is not obvious what those achievements might be. I have already argued in Chapter 7 that moral properties and facts do not seem to play any ineliminable role in the best explanations of whatever practical successes we have achieved utilizing that discourse. On the face of it, this finding would seem to imply that, to the extent that our moral terms are causally regulated by certain natural properties, the fact that these terms are causally regulated in this way does not help to explain any theoretical or practical successes achieved using them. If so, then it follows from Boyd's updated semantics that the central terms of our moral discourse fail to refer. In that case, moral nihilism is true and moral realism is false.

[9] Several examples of this kind of predictive success were discussed in §6.5.2.

A.3.2. Practical success as facilitating well-being.

Perhaps the argument of the previous section is too hasty. Although Boyd acknowledges that "it is a controversial issue just what sorts of successes would count as *moral* successes," he suggests, nevertheless that what we achieve (or what we can achieve) through our use of moral discourse is "the well-being of people generally" (Boyd 2003a: 516f, 2003b: 36, emphasis in the original). Now, I count myself among those who are skeptical that such an achievement, if it is real, is best viewed as a success of *moral* theory, rather than, say, psychological theory.[10] But let us waive this skepticism and suppose that moral discourse really does play an indispensable role in our achieving the well-being of people. With this supposition granted, those who would press the MTE argument against Boyd's semantics must make plausible the further stipulations that (a) Twin Earthlings are able to successfully achieve some measure of *well-being* through employing terms like t-'right' etc., and (b) this success is partly explained by the fact that T^d properties causally regulate the use of their terms like t-'right' etc.

What are the prospects for the anti-naturalist defender of MTE with respect to meeting this challenge? Can an MTE scenario be described that coherently incorporates these two additional stipulations?

One thing that makes answering these questions difficult is that the question of what constitutes *well-being* itself depends upon the results of first-order theories in axiology. Presumably, then, the term 'well-being' as used on Twin Earth is causally regulated by a different natural property than the natural property that regulates its use on

[10] Provided that I have a fairly clear idea of the sort of mental or physical condition that I want to be in (i.e., of what sort of property I view as *flourishing*-making), I do not seek out the writings of philosophers for advice on how to achieve this condition; for that, I address my inquiries to psychologists, nutritionists, and physical trainers. I expect things would be no different if I were concerned to discover how best to bring about this condition for a multitude of people in my community, rather than just myself.

Earth. This raises the question of *whose* standard of *well-being* we should use to determine whether Twin Earthlings have succeeded in achieving some level of flourishing. There appears to be three options: First, Twin Earthlings' success in achieving *well-being* should be judged by their own Twin Earthling standards (i.e., by the standards prescribed by the theory T^d, which specifies the essences of the properties to which Twin moral terms putatively refer). Second, we should judge Twin Earthlings' successes using our own Earthling axiological standards; that is, we credit Twin Earthling individuals with achieving a measure of flourishing when they exemplify the natural property that causally regulates our own Earthling uses of 'well-being' and 'flourishing,' i.e., a property specified by T^c. And the final option is that we should judge their and our own successes by some very broad axiological theory that is somehow compatible with both T^d and T^c.

Now, it is not entirely clear to me how the third option can be worked out. Presumably, the axiologies associated with T^d and T^c respectively will give divergent verdicts about the welfare-value of at least some lives. If they do, it is hard to see how to formulate a broad, substantive axiology that is compatible with both evaluative theories. For this reason, I recommend that we set option three aside. Let's consider, instead, the first option. According to this, we are to judge Twin Earthlings' practical successes using the axiological standards enshrined in their own moral theory, T^d. Under this constraint, I see no reason why it is impossible that there be a scenario where Twin Earthling discourse satisfies Boyd's achievement explanation condition. Presumably, *well-being* for a Twin Earthling consists in living a life in accordance with some standard prescribed by T^d. This standard specifies what natural properties must be instantiated by a life in

order for it to count as flourishing for Twin Earthlings. If, in fact, these natural properties are what causally regulate Twin uses of t-'good,' then it would seem that Twin Earthlings can come to better know the relevant axiological standard by investigating which natural properties regulate their uses of this predicate. In turn, this knowledge would seem to facilitate their success at satisfying this axiological standard, i.e., their success at living good, flourishing lives. To put all of this together: The causal regulation of t-'good' by these natural properties explains Twin Earthlings' knowledge of their own axiological standard; this knowledge explains their ability to satisfy this standard, and thus, to successfully achieve flourishing. Assuming the transitivity of explanation, then, we may conclude that the causal regulation of twin moral discourse by the relevant natural properties explains the practical successes achieved by Twin Earthlings through utilizing that moral discourse. If all of this may be coherently conceived, then there is no problem describing a Twin Earth scenario in which Twin Earthlings satisfy *both* Boyd's causal regulation condition and the achievement explanation condition for reference. The addition of the achievement explanation condition to Boyd's semantics does not threaten the cogency of the MTE argument.

It might be argued, however, that when judging the practical successes of twin moral discourse, we should do so utilizing our own Earthling standard of human flourishing, as captured by T^c. This is the second option mentioned above. Hitherto, I have not specified what sort of axiology is incorporated into T^c. Let's suppose that T^c includes a simple hedonist theory of *well-being*. (According to this theory, the well-being of a life is a measure of the amount of pleasure the life contains, minus the amount of pain it contains.) By contrast, we should suppose that the axiology of T^d is a non-

hedonistic theory (perhaps it is a form of perfectionism). Even with these stipulations, I see no reason to think that a coherent MTE scenario cannot be told according to which the causal regulation of twin moral discourse by T^d properties nevertheless explains the Twins' practical success at achieving a high level of flourishing, where this flourishing is measured by the hedonistic axiology (which, we are supposing, the Twin Earthlings do not themselves accept). This could be explained by Twin Earthlings' different psychological temperament. Due to their temperament, the Twin Earthlings best achieve *well-being* by hedonist standards when they collectively subscribe to, and act in accordance with, their deontological theory of *rightness* and perfectionist theory of *value*. In this way, twin moral discourse could satisfy the achievement explanation condition even when we measure their success by our own Earthling standard of *well-being*.[11]

A.3.3. Predictive success.

It may be worth considering one more example of putative practical success that is arguably facilitated by moral discourse. In Chapter 7, I discussed several examples of empirical predictions generated by moral theories. As I suggested in §A.3.1 above, I believe that the arguments I advance in that chapter warrant the conclusion that, if natural properties do causally regulate the use of our moral terms, this fact is not part of the best explanation of the successful predictions we make using those terms and our best moral

[11] Against this, the defender of SEN may insist that, in the case I have described, T^c properties rather than T^d properties are what are regulating the Twins' use of moral terms, even if the Twins are unaware of this. I do not know exactly how this objection would proceed, but I see a temptation to make it. In any event, I think it can be sidestepped. I see no reason to insist that Twin Earthlings' use of moral terms succeed in referring to T^d properties only if the regulation between their terms and the T^d properties help to explain the Twins' ability to achieve *well-being* by our T^c standards. For if we did insist on such a condition, we would be forced to conclude that, where the regulation of moral terms by T^d properties fails to result in and explain the Twins' success at achieving *well-being* by our Earthling standards, twin moral terms simply fail to refer, even if they achieve success by their own standards. It seems to me that this sort of construal of the MTE scenario is unmotivated, at least given realist assumptions about moral discourse.

theories. But suppose I am wrong about this. Let us revisit the example involving empirical predictions based on judgments about an agent's moral character and consider its implications for MTE, given Boyd's achievement explanation condition. The question before us is this: can we coherently imagine that Twin Earthlings are able to make successful predictions of this sort utilizing their own moral vocabulary and moral theories (while, at the same time, Earthlings are also able to achieve a similar success using their moral vocabulary and moral theories)? I see no reason to think not. Horgan and Timmons' have already stipulated that Twin Earthlings tend to perform those actions that they judge to be "right." Since we know that the sorts of actions that they judge to be right are just those actions that treat no one as a mere means, and since, presumably, Twin Earthlings apply t-'good' roughly to all and only those agents who have an especially strong tendency to perform only actions that treat no one as a mere means, Twin Earthlings should be able to make successful predictions about the behavior of those to whom t-'good' is properly applied. For example, suppose that some Twin Earthlings properly judge that t-'good' applies to Dr. Smith. And suppose, further, that they properly judge that t-'wrong' applies to any act of organ harvesting, because such acts involve treating someone as a mere means. Given these facts, we should expect Twin Earthlings will be able to successfully predict that Dr. Smith will not perform the act of organ harvesting.[12] If so, then, again, there seems to be no reason to deny that a coherent MTE scenario can be described in which Boyd's achievement explanation condition for reference is satisfied along with his causal regulation condition.

[12] Given the sorts of complaints I raise above in §7.4.2, this may be false. But again, I am waiving those complaints here in order to make favorable assumptions on behalf of naturalistic moral realism.

Perhaps there are further examples of achievements that are won through the use of moral discourse that yield more favorable results for naturalistic moral realists with respect to the present concern. If so, I do not know what those examples are. Until such examples are produced, I believe that the above considerations license us in concluding for now that that Boyd's amended semantics poses no special problem for the Moral Twin Earth argument.

A.4. Partial denotation.

H&T formulate Boyd's CSN in such a way that it requires a natural property to *uniquely* causally regulate the use of a moral term t in order for that property to be designated by t. However, in his (1988: 226) Boyd leaves open the possibility that more than one natural property causally regulates our use of 'right.' In a scenario of this kind, according to Boyd, the term 'rightness' would *partially* denote each of those properties.[13] A plausible illustration of partial denotation is provided by the kind term 'jade.' It is often claimed that 'Jade' partially denotes the mineral jadeite and partially denotes the mineral nephrite.[14] Given Boyd's semantics, the explanation for this is presumably that our use of 'jade' is causally regulated by both kinds of mineral. (In addition, we may have to add that neither mineral better explains the achievements of 'jade'-talk than the other).

With respect to the MTE scenario, an appeal to partial denotation might be thought to furnish a way to provide a common referent for 'rightness' and t-'rightness':

[13] Here I follow Boyd in using the term 'denotes' to express the semantic relationship between a term and a kind or property. While some philosophers draw a distinction between the semantic relations expressed by 'denotes,' 'refers,' and 'designates,' others seem to use these terms interchangeably. For the purposes of this appendix, I follow the latter practice.

[14] As far as I know, this claim entered the philosophical literature through Putnam (1975b: 241). I have seen it repeated in a number of commentaries on Putnam, usually without objection.

just as ‘jade’ partially denotes jadeite and partially denotes nephrite, it might be argued that ‘rightness’ and t-‘rightness’ both partially denote both *maximizing utility* and *treating no one as a mere means*. If so, these two terms have a common denotation; moreover, their corresponding predicates (‘right’ and t-‘right’) express a common content. Because of this, it might be argued, CSN does not entail a form of conceptual relativism with respect to the MTE scenario.

The partial denotation reply will note work. The trouble becomes apparent when we consider what we should say about the truth-conditions of sentences in which partially denoting terms occur. To see the trouble, let’s return to the example involving ‘jade.’ Consider, first, a relatively unproblematic sentence: ‘Jade is a mineral.’ This sentence attributes to jade a property that, as it happens, is instantiated by *both* jadeite and nephrite. Because of this, there seems to be no difficulty in holding that ‘Jade is a mineral’ expresses a truth. The more difficult case involves a sentence such as ‘Jade is partly composed of aluminum.’ If ‘jade’ uniquely denoted jadeite, this sentence would express a truth, since jadeite is partly composed of aluminum. If ‘jade’ uniquely denoted nephrite, this sentence would express a falsehood, since nephrite is not partly composed of aluminum. However, as things actually are, ‘jade’ does not uniquely denote either of these minerals; rather, it partially denotes both. What, then, is the truth-value of ‘Jade is partly composed of aluminum’? There seems to be two salient options. First, we can say that the truth-value of this sentence is indeterminate. Second, we could say that, relative to one disambiguation, the sentence is true, and relative to another disambiguation, it is false (cf. Field 1973).

Assuming that 'rightness' and t-'rightness' partially denote *maximizing utility* and *treating no one as a mere means*, the same interpretive decision is forced on us with respect to certain moral sentences. Consider, for example, the question of what truth-value we should assign to the sentence 'the organ harvest instantiates rightness.' Given that the organ harvest has the property of *maximizing utility*, but lacks the property of *treating no one as a mere means*, and given supposition that 'rightness' partially denotes both of these properties, it seems that we are forced to say either that the sentence in question has an indeterminate truth-value, or else that it is true according to one disambiguation of 'rightness,' but false with respect to another.

Whichever of these options we choose, the prospects for building a satisfying defense of SEN against MTE are dim. In the first place, the reason for supposing that 'jade' partially denotes jadeite and nephrite (at least, assuming the truth of Boyd's causal regulation semantics), is that both of these minerals causally regulate our uses of 'jade.' If only one of these minerals had causally regulated the use of 'jade,' it would be wrong to say that 'jade' partially denotes both. Notice, however, that in the MTE story, it is stipulated that only one property (*maximizing utility*) causally regulates the use of 'rightness' while a different, but unique, property (treating *no one as a mere means*) causally regulates the use of t-'rightness.' Given this stipulation, there is simply no justification for supposing that 'rightness' and t-'rightness' both partially denote the same properties. To suggest this is to ignore the way the scenario is described. Consequently, the observation that 'moral rightness' might partially denote different properties on our planet in the actual world does nothing to supply the naturalist moral realist with a reply to MTE.

At best, the appeal to partial denotation might be employed to answer concerns about a moral term being regulated by multiple distinct natural properties within a single linguistic community. But even here it does no good. If our use of 'rightness' is causally regulated by multiple distinct natural properties, then in any circumstance in which these properties fail to be coextensive, an attribution of moral rightness will be indeterminate. If, for example, 'right' were causally regulated by both T^c and T^d properties, it would follow that, prior to any disambiguation, there is simply no fact of the matter as to whether (e.g.) it is right to keep our promises when the utility of breaking them is greater. This seems to me to be a problem: while it is certainly unfair to require of moral realism that *every* act-token have a determinate moral status, this sort of indeterminacy is just too easy to stumble upon. The result would be that in all cases where consequentialists and deontologists reasonably disagree about the moral status of an act (where this disagreement is not due to ignorance of the non-moral facts), there would be no determinate answer as to who, if anyone, is correct. Surely this is too much indeterminacy.

I have suggested that, under the present assumptions, the partial denotation of moral terms would render too many moral utterances indeterminate with respect to their truth-values. To avoid this result, we might insist that partially denoting moral terms that appear in utterances must be disambiguated before we attempt to assign truth-values to those utterances. This maneuver permits us to say of the sentence, 'Breaking promises to maximize utility instantiates rightness,' that it is true according to one disambiguation of 'rightness,' but false according to another disambiguation. Thus, there is no need to posit rampant indeterminacy.

The trouble with the disambiguation approach is that it leaves it unclear which disambiguation of 'rightness' to use in which context. If there are no general rules to guide us, then partisans of deontology may simply insist that we always understand their uses of 'rightness' to denote *treating no one as a mere means*; meanwhile partisans of consequentialism might insist that we always understand their uses of 'rightness' to denote *maximizing utility*. But this is surely unacceptable; for it implies the most vulgar kind of moral subjectivism. But suppose that we are optimistic that rules for disambiguation can be found. What might those rules be? With respect to utterances of 'jade' it is natural to suppose that the principle of charity dictates which disambiguation to select. That is to say, we should select the interpretation of 'jade' that makes more of the speaker's beliefs come out true. But notice that, if we try to apply this principle to uses of 'rightness,' we will very likely be plunged back into moral subjectivism. For if someone subscribes to T^d, then surely the way to make most of her utterances involving 'rightness' to come out true is to suppose that the term denotes *treating no one as a mere means*. Likewise, if someone subscribes to T^c, then the way to make most of his utterances of 'rightness' come out true is by taking the term to denote *maximizing utility*. This kind of moral subjectivism, whereby the truth-conditions of a person's moral utterances depend upon whichever moral standard she happens to subscribe to, almost certainly violates moral realism's stance independence requirement; and so, it would seem to entail the falsity of moral realism. But whatever the case may be, this form of subjectivism is surely an unappealing metaethical commitment in its own right. I conclude then, that the adoption of a partial denotation maneuver cannot be used to rescue naturalistic moral realism from the MTE argument.

BIBLIOGRAPHY

Adams, Robert. 1979. "Divine Command Metaethics Modified Again," *Journal of Religious Ethics* 7: 66-79.

Armstrong, D. M. 1978. *Universals and Scientific Realism Volume II: A Theory of Universals*, Cambridge: Cambridge University Press.

Ayer, Alfred Jules. 1936/1952. *Language, Truth and Logic*, New York: Dover.

Bealer, George. 2000. "A Theory of the A Priori," *Pacific Philosophical Quarterly* 81: 1-30.

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

_____. 1988. "How to be an Ethical Anti-Realist," *Midwest Studies in Philosophy* 12: 361-375

_____. 1998. *Ruling Passions*, New York: Oxford University Press.

Boyd, Richard. 1973. "Realism, Underdetermination, and a Causal Theory of Evidence," *Nous*, 7(1): 1-12.

_____. 1979. "Metaphor and Theory Change," in Andrew Ortony (ed.), *Metaphor and Thought*, Cambridge: Cambridge University Press.

_____. 1980. "Materialism Without Reductionism: What Physicalism Does Not Entail," In Ned Block (ed.), *Readings in the Philosophy of Psychology*, Cambridge: Harvard University Press: 67-106.

_____. 1982. "Scientific Realism and Naturalistic Epistemology," In P. D. Asquith and R. N. Giere (eds.), *PSA 1980*, Vol. 2. East Lansing: Philosophy of Science Association.

_____. 1983. "On the Current Status of the Issue of Scientific Realism," *Erkenntnis* 19: 45-90.

_____. 1988. "How to be a Moral Realist," in Sayre-McCord, Geoffrey (ed.) *Essays on Moral Realism*, Ithaca: Cornell University Press.

_____. 1990. "Realism, Approximate Truth, and Philosophical Method," Reprinted in David Papineau (ed.), *The Philosophy of Science*. Oxford: Oxford University Press (1996): 215-255.

_____. 1999a. "Homeostasis, Species, and Higher Taxa," in Robert A. Wilson (ed.), *Species: New Interdisciplinary Essays*, Cambridge: MIT Press.

_____. 1999b. "Kinds, Complexity and Multiple Realization: Comments on Millikan's 'Historical Kinds and the Special Sciences'," *Philosophical Studies* 95: 67-98.

_____. 1999c. "Kinds as the "Workmanship of Men": Realism, Constructivism, and Natural Kinds," in Julian Nida-Rümelin (ed.), *Rationalität, Realismus, Revision: Proceedings of the Third International Congress, Gesellshaft für Analytische Philosophie*, Berlin: de Gruyter.

_____. 2003a. "Finite Beings, Finite Goods: The Semantics, Metaphysics, and Ethics of Naturalist Consequentialism, Part I," *Philosophy and Phenomenological Research* 66(3), 505-553.

_____. 2003b. "Finite Beings, Finite Goods: The Semantics, Metaphysics, and Ethics of Naturalist Consequentialism, Part II," *Philosophy and Phenomenological Research* 67(1), 24-47.

Brandt, Richard B. 1959. *Ethical Theory*, Englewood Cliffs, NJ: Prentice-Hall.

Brink, David O. 1984. "Moral Realism and the Sceptical Arguments from Disagreement and Queerness," *Australasian Journal of Philosophy* 62: 111-25.

_____. 1986. "The Status of Morality," *The Philosophical Review* 95(1): 144-147.

_____. 1987. "Ethics, Persuasion and Truth," *The Philosophical Review* 96(2): 290-295.

_____. 1989. *Moral Realism and the Foundation of Ethics*, Cambridge: Cambridge University Press.

_____. 1999. "Objectivity and Dialectical Methods in Ethics," *Inquiry* 42(2): 195-212

_____. 2001. "Realism, Naturalism, and Moral Semantics," *Social Philosophy and Policy* 18(2): 154-176.

Broome, John. 2004. "Reasons," in Wallace, Pettit, Scheffler, and Smith (eds.), *Reason and Value*, Oxford: Clarendon Press.

Carnap, Rudolph. 1947/1956. *Meaning and Necessity*, Chicago: University of Chicago Press, (1988).

_____. 1955/1956. "Meaning and Synonymy in Natural Languages," reprinted in *Meaning and Necessity*, Chicago: University of Chicago Press (1988).

Chalmers, David J. (1996). *The Conscious Mind*, Oxford: Oxford University Press.

Copp, David. 1990. "Explanation and Justification in Ethics," *Ethics*, 100(2): 237-258.

_____. 1995. *Morality, Normativity, and Society*, Oxford: Oxford University Press.

_____. 2000. "Milk, Honey, and the Good Life on Moral Twin Earth," *Synthese* 124: 113-137.

_____. 2003. "Why Naturalism?" *Ethical Theory and Moral Practice* 6: 179-200.

_____. 2005. "A Skeptical Challenge to Moral Non-Naturalism and a Defense of Constructivist Naturalism," *Philosophical Studies* 126(2): 269-283.

Dancy, Jonathan. 1993. *Moral Reasons*, Cambridge, MA: Blackwell Publishers.

Daniels, Norman. 1979. "Wide Reflective Equilibrium and Theory Acceptance in Ethics," *The Journal of Philosophy* 76(5): 256-282.

Davidson, Donald. 1963. "Actions, Reasons, and Causes," *The Journal of Philosophy* 60(23): 685-700.

Devitt, Michael. 2005. "Rigid Application," *Philosophical Studies* 125: 139-165.

de Waal, Frans. 1982. *Chimpanzee Politics*, Baltimore: Johns Hopkins University Press.

Donnellan, Keith S. 1970. "Proper Names and Identifying Descriptions," *Synthese* 21: 335-358.

_____. 1983. "Kripke & Putnam on Natural Kind Terms," In Ginet, Carl, and Sydney Shoemaker (eds.), *Knowledge and Mind: Philosophical Essays*, New York: Oxford University Press: 84-104.

Doris, John M. 2002. *Lack of Character*, New York: Cambridge University Press.

Duhem, Pierre. 1906/1914. *The Aim and Structure of Physical Theory, 2nd Edition*. Princeton, NJ: Princeton University Press (1954).

Dupré, John. 1981. "Natural Kinds and Biological Taxa," *The Philosophical Review* 90(1): 66-90.

Ellis, Brian. 2001. *Scientific Essentialism*, New York: Cambridge University Press.

Epstein, Seymour, and Edward J. O'Brien. 1985. "The Person-Situation Debate in Historical and Current Perspective," *Psychological Bulletin* 98(3): 513-537.

Evans, Gareth. 1973. "The Causal Theory of Names," *Aristotelian Society*, Suppl. 47: 187-208

Ewing, A. C. 1953. *Ethics*, New York: Macmillan Publishing Co.

Feldman, Fred. 1978. *Introductory Ethics*, Upper Saddle River, NJ: Prentice Hall.

_____. 2005. "The Open Question Argument: What it Isn't; and What it Is," *Philosophical Issues* 15: 22-43.

Field, Hartry. 1973. "Theory Change and The Indeterminacy of Reference," *The Journal of Philosophy*, 70(14): 462-481.

Firth, Roderick. 1952. "Ethical Absolutism and the Ideal Observer," *Philosophy and Phenomenological Research* 12(3): 317-345.

Foot, Philippa. 1958. 'Moral Arguments,' *Mind* 67: 502-513.

Freud, Sigmund. 1931/1961. *Civilization and Its Discontents*, New York: Norton.

_____. 1933/1965. *New Introductory Lectures on Psycho-Analysis*, New York: Norton.

Gample, Eric H. 1997. 'Ethics, Reference, and Natural Kinds,' *Philosophical Papers* 26(2): 147-163.

Geirsson, Heimer. 2003. 'Moral Twin Earth: The Intuitive Argument,' *Southwest Philosophy Review* 19(1): 115-124.

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

_____. 1965. "Assertion," The Philosophical Review 74(4): 449-465.

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

Haidt, Jonathan. 2001. "The Emotional Dog and Its Rational Tail," *Psychological Review* 108(4): 814-834.

Hamilton, W. D. 1964a. "The Genetical Evolution of Social Behavior. I," *Journal of Theoretical Biology* 7: 1-16.

_____. 1964b. "The Genetical Evolution of Social Behavior. II," *Journal of Theoretical Biology* 7: 17-52.

Haney, Craig., Curtis W. Banks, and Philip G. Zimbardo. 1973. "Study of Prisoners and Guards in a Simulated Prison," *Naval Research Reviews* 9: 1–17.

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

_____. 1963. *Freedom and Reason*, New York: Oxford University Press.

Harman, Gilbert. 1975. "Moral Relativism Defended," *The Philosophical Review* 84(1): 3-22.

_____. 1977. *The Nature of Morality*, New York: Oxford University Press.

_____. 1985. "Is There a Single True Morality?" in David Copp and David Zimmerman (eds.), *Morality Reason and Truth*, Totowa, NJ: Rowman & Littlefield.

Huemer, Michael. 2005. *Ethical Intuition*, New York: Palgrave Macmillan.

Horgan, Terence and Mark Timmons. 1990-91. "New Wave Moral Realism Meets Moral Twin Earth," *Journal of Philosophical Research* 16: 447-465.

_____. 1992a. "Troubles on Moral Twin Earth: Moral Queerness Revived," *Synthese* 92: 221-260.

_____. 1992b. "Troubles for New Wave Moral Semantics: The Open Question Argument Revived," *Philosophical Papers* 21: 153-175.

_____. 1996a. "From Moral Realism to Moral Relativism in One Easy Step," *Critica* 28: 3-39.

_____. 1996b. "Troubles for Michael Smith's Metaethical Rationalism," *Philosophical Papers* 25(3): 203-231.

_____. 2000. "Copping Out on Moral Twin Earth," *Synthese* 124: 139-152.

_____. Forthcoming. "Analytical Moral Functionalism Meets Moral Twin Earth," in I. Ravenscroft (ed.), *Essays on the Philosophy of Frank Jackson*, Oxford: Oxford University Press.

Hilpinen, Risto. 1976. "Approximate Truth and Truthlikeness," In Marian Przelecki, Klemens Szaniawski, and Ryszard Wojcicki (eds.), *Formal Methods in the Methodology of Empirical Sciences*, Boston: Reidel

Hume, David. 1739/2000. *A Treatise of Human Nature*, New York: Oxford University Press.

Isen, Alice, and Paula Levin. 1972. "Effect of Feeling Good on Helping: Cookies and Kindness," *Journal of Personality and Social Psychology* 21(3): 384-388.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford: Oxford University Press.

Jackson, Frank, and Philip Pettit. 1988. "Functionalism and Broad Content," *Mind* 97(387): 381-400.

Joyce, Richard. 2001. *The Myth of Morality*, Cambridge: Cambridge University Press.

_____. 2006. *The Evolution of Morality*, Cambridge, MA: The MIT Press.

Kagan, Shelly. 1998. "Rethinking Intrinsic Value," *Journal of Ethics* 2(4): 277-297.

_____. 1998. *Normative Ethics*, Boulder, CO: Westview Press.

Kant, Immanuel. 1783/2004. *Prolegomena to any Future Metaphysics*, New York: Oxford University Press.

Kitcher, Philip. 1984. "Species," *Philosophy of Science* 51(2): 308-333.

_____. 1998. "Psychological Altruism, Evolutionary Origins, and Moral Rules," *Philosophical Studies* 89: 283: 316.

_____. 2001. "Real Realism: The Galilean Strategy," *The Philosophical Review* 110(2): 151-197

_____. 2006. "Biology and Ethics," in David Copp (ed.), *The Oxford Handbook of Ethical Theory*, Oxford: Oxford University Press.

Korsgaard, Christine. 1986. "Skepticism about Practical Reason," *The Journal of Philosophy* 83(1): 5-25.

Kornblith, Hilary. 1993. *Inductive Inference and Its Natural Ground*, Cambridge: MIT Press.

_____. 1994. "Naturalism: Both Metaphysical and Epistemological," Midwest Studies in Philosophy 19: 39-52.

_____. 2002. *Knowledge and Its Place in Nature*, New York: Oxford University Press.

Kraemer, Eric R. 1990-91. "On the Moral Twin Earth Challenge to New Wave Moral Realism," *Journal of Philosophical Research* 16: 467-472.

Kripke, Saul. 1971. "Identity and Necessity," reprinted in Stephen P. Schwartz (ed.), *Naming, Necessity, and Natural Kinds*, Ithaca: Cornell University Press (1977): 67-101.

_____. 1980. *Naming and Necessity*, Cambridge, MA: Harvard University Press.

Kuhn, Thomas S. 1962/1996. *The Structure of Scientific Revolutions, Third Edition*. Chicago: University of Chicago Press.

_____. 1969/1996. "Postscript," In *The Structure of Scientific Revolutions, Third Edition*, Chicago: University of Chicago Press.

LaPorte, Joseph. 2004. *Natural Kinds and Conceptual Change*, Cambridge: Cambridge University Press.

Laurence, Stephen, Eric Margolis, and Angus Dawson. 1999. 'Moral Realism and Twin Earth,' *Facta Philosophica* 1: 135-165.

Leiter, Brian. 2001. "Moral Facts and Best Explanations," *Social Philosophy and Policy* 18(2): 79-101.

Lenman, James. 1999. "The Externalist and the Amoralist." *Philosophia* 27: 441-457.

_____. 2000. "Consequentialism and Cluelessness," Philosophy & Public Affairs 29(4): 343-370.

Leplin, Jarrett. 1997. *A Novel Defense of Scientific Realism*, New York: Oxford University Press.

Lewis, David. 1970. "How to Define Theoretical Terms," *Journal of Philosophy* 67(16): 427-446.

_____. 1980. "Mad Pain and Martian Pain," reprinted in *Philosophical Papers Vol. I,* New York: Oxford University Press (1983): 122-132.

_____. 1994. "Reduction of Mind," Reprinted in *Papers in Metaphysics and Epistemology*, Cambridge, Cambridge University Press (1999): 291-324.

Lycan, W.G. 1988. *Judgment and Justification*, Cambridge: Cambridge University Press.

Lyons, David. 1976. "Ethical Relativism and the Problem of Incoherence," *Ethics* 86(2): 107-121.

Mackie, John L. 1977. *Ethics: Inventing Right and Wrong*, New York: Penguin Books.

Mallon, Ron. 2003. "Social Construction, Social Roles, and Stability," in Frederick Schmitt (ed.), *Socializing Metaphysics: The Nature of Social Reality*, New York: Rowman and Littlefield: 327-353.

Marx, Karl, and Friedrich Engels. 1933/1978. *The German Ideology*, in Robert Tucker (ed.), *The Marx-Engels Reader* (2nd edition), New York: Norton.

Mayr, Ernst. 1996. "What is a Species, and What is Not?" *Philosophy of Science* 63(2): 262-277.

McDowell, John. 1981/1998. "Non-Cognitivism and Rule-Following," reprinted *in Mind, Value, and Reality*, Cambridge, MA: Harvard University Press.

Mellor, D. H. 1977. 'Natural Kinds,' *The British Journal for the Philosophy of Science* 28(4): 299-312.

Merli, David 2002. "Return to Moral Twin Earth," *Canadian Journal of Philosophy* 32(2): 207-240.

Milgram, Stanley. 1963. "Behavioral Study of Obedience," *Journal of Abnormal and Social Psychology* 67(4): 371-378.

Mill, John Stuart. 1867. *A System of Logic*, New York: Harper and Brothers Publishers.

Miller, Richard. 2000. "Half-Naturalized Social Kinds," *Philosophy of Science* 67, Supplement: 640-652.

Millikan, Ruth Garrett. 2000. *On Clear and Confused Ideas*, Cambridge: Cambridge University Press.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press (1993).

"Moral." *The American Heritage Dictionary of the English Language,* 4th ed. Boston: Houghton Mifflin, 2000. www.bartleby.com/61/. (Nov. 8, 2006).

"Moral." *Merriam-Webster Online Dictionary*. 2006. http://www.merriam-webster.com (Nov. 8, 2006).

"Moral." *Oxford English Dictionary Online*. 2006. http://dictionary.oed.com./ (Nov. 8, 2006).

Musgrave, Alan. 1988. "The Ultimate Argument for Scientific Realism," In Nola, Robert (ed.), *Relativism and Realism in Science*: 229-252.

Nagel, Thomas. 1970. *The Possibility of Altruism*, Princeton: Princeton University Press.

_____. 1986. *The View from Nowhere*, New York: Oxford University Press.

Nichols, Shaun. 2002. "Norms With Feeling: Towards a Psychological Account of Moral Judgment," *Cognition* 84: 221-236.

_____. 2004. *Sentimental Rules*, New York: Oxford University Press.

Nietzsche, Friedrich. 1887/1967. *On The Genealogy of Morals*, translated by Kaufman, Walter and R. J. Hollingdale, New York: Vintage.

Nisbett, Richard, and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, NJ: Prentice-Hall.

_____. 1991. *The Person and the Situation*, New York: McGraw-Hill.

Nisbett, Richard and Dov Cohen. 1996. *Culture of Honor: The Psychology of Violence in the South*, Boulder, CO: Westview Press.

Nolan, Daniel, Gregg Restall, and Caroline West. 2005. "Moral Fictionalism Versus the Rest," *Australasian Journal of Philosophy* 83(3): 307-330.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*, New York: Basic Books.

Oddie, Graham. (1999). "Moral Realism, Moral Relativism and Moral Rules," *Synthese* 117: 241-274.

Plato. 1974. *The Republic*, translated by G. M. A. Grube, Indianapolis: Hackett.

Putnam, Hilary. 1967. "The Nature of Mental States," reprinted in *Mind, Language and Reality*, Cambridge: Cambridge University Press (1975): 429-440.

_____. 1974. "The Corroboration of Theories," reprinted in *Mathematics, Matter and Method*, Cambridge: Cambridge University Press (1975): 250-269.

_____. 1975a. "Language and Reality," in *Mind, Language and Reality*, New York: Cambridge University Press (1975): 272-290.

_____. 1975b. "The Meaning of 'Meaning'," reprinted in *Mind, Language and Reality*, New York: Cambridge University Press (1975): 215-271.

_____. 1975c. "What is Mathematical Truth?" reprinted in *Mathematics, Matter and Method*. Cambridge: Cambridge University Press (1975): 60-78.

_____. 1977. "Realism and Reason," *Proceedings and Addresses of the American Philosophical Association* 50(6): 483-498.

_____. 1981. *Reason, Truth and History*, New York: Cambridge University Press.

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

_____. 1969. "Natural Kinds," In *Ontological Relativity and Other Essays*, New York: Columbia University Press.

Quinn, Warren S. 1986. "Truth and Explanation in Ethics," *Ethics* 96(3): 524-544.

Railton, Peter. 1986. "Moral Realism," *Philosophical-Review* 95: 163-207.

_____. 1989. 'Naturalism and Prescriptivity,' *Social Philosophy and Policy* 7(1): 151-174.

_____. 1993. 'Noncognitivism about Rationality: Benefits, Costs, and an Alternative,' *Philosophical Issues* 4: 36-51.

_____. 1995. "Subject-ive and Objective," *Ratio* 8(3): 259-276.

_____. 1996. "Moral Realism: Prospects and Problems," in Walter Sinnott-Armstrong and Mark Timmons (eds.), *Moral Knowledge?* New York: Oxford University Press: 49-81.

_____. 2000. "Darwinian Building Blocks," *Journal of Consciousness Studies* 7(1-2): 55-60.

Rand, Ayn. 1971/1999. *Return of the Primitive: The Anti-Industrial Revolution*, New York: Penguin Putnam.

Rawls, John. 1951. "Outline of a Decision Procedure for Ethics," *The Philosophical Review* 60(2): 177-197.

_____. 1971/1999 *A Theory of Justice* (Revised Edition), Cambridge, MA: Harvard University Press.

_____. 1980. "Kantian Constructivism in Moral Theory," *The Journal of Philosophy* 77(9): 515-572.

Ridge, Michael. 2008. "Moral Non-Naturalism," *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/fall2008/entries/moral-non-naturalism/>.

Ross, W. D. 1930. *The Right and the Good*, Indianapolis: Hackett.

Ruse, Michael (1986): *Taking Darwin Seriously: A Naturalistic Approach to Philosophy*, Amherst, NY: Prometheus Books (1998).

Russell, Bertrand. 1948. *Human Knowledge: Its Scope and Limits*, New York: Simon and Schuster.

Sayre-McCord, Geoffrey. 1988. "Moral Theory and Explanatory Impotence," reprinted in *Essays on Moral Realism*, Ithaca, New York: Cornell University Press.

_____. 1991. "Being a Realist About Relativism," *Philosophical Studies* 61: 155-176.

Shafer-Landau, Russ. 2003. *Moral Realism*, Oxford: Oxford University Press.

_____. 2006. "Ethics as Philosophy: A Defense of Ethical Nonnaturalism." In Horgan, Terry, and Mark Timmons (eds.), *Metaethics after Moore*, New York: Oxford University Press: 209-232.

Shoemaker, Sydney. 1981. "Some Varieties of Functionalism," reprinted in *Identity, Cause, and Mind* (2nd Edition), New York: Oxford University Press (2003): 261-268.

Sidgwick, Henry. 1907. *The Methods of Ethics*, Indianapolis: Hackett (1981).

Smart, J. J. C. 1963. *Philosophy and Scientific Realism*, New York: Routledge & Kegan Paul

Smith, Michael. 1994. *The Moral Problem*, Oxford: Basil Blackwell.

_____. 1995. "Internal Reasons," *Philosophy and Phenomenological Research* 55(1): 109-131.

Sociobiology Study Group. 1977. "Sociobiology," In the Ann Arbor Science for the People Editorial Collective (ed.), *Biology as a Social Weapon*, Minneapolis: Burgess Publishing Co.

Sterelny, K. .1983. "Natural Kind Terms," *Pacific Philosophical Quarterly* 64: 110-125.

Stevenson, Charles L. 1937. "The Emotive Meaning of Ethical Terms," *Mind* 46 (181): 14-31.

_____. 1944. *Ethics and Language*, New Haven, CT: Yale University Press.

Street, Sharon. 2006. "A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies* 127: 109-166.

Sturgeon, Nicholas L. 1984. "Moral Relativism; Virtues and Vices, and Other Essays in Moral Philosophy," *The Journal of Philosophy* 81(6): 325-333.

_____. 1985a. "Moral Explanations," In David Copp and David Zimmerman (eds.), *Morality Reason and Truth*. Totowa, NJ: Rowman & Littlefield.

_____. 1985b. "Gibbard on Moral Judgment and Norms," *Ethics* 96(1): 22-33.

_____. 1986a. "Harman on Moral Explanations of Natural Facts," *The Southern Journal of Philosophy* 24, Supplement: 69-78.

_____. 1986b. "What Difference Does It Make Whether Moral Realism is True?" *The Southern Journal of Philosophy* 24, Supplement: 115-141.

_____. 1991. "Contents and Causes: A Reply to Blackburn," *Philosophical Studies* 61: 19-37.

_____. 1994. "Moral Disagreement and Moral Relativism," *Social Philosophy and Policy*. 11(1): 80-115.

_____. 2002. "Ethical Intuitionism and Ethical Naturalism," In Philip Stratton-Lake (ed.), *Ethical Intuitionism: Re-evaluations*, New York: Oxford University Press.

_____. 2003. "Moore on Ethical Naturalism," *Ethics* 113(3): 528-556.

_____. 2006a. "Moral Explanations Defended," in James Dreier (ed.), *Contemporary Debates in Moral Theory*, Malden, MA: Blackwell Publishing.

_____. 2006b. "Ethical Naturalism," in David Copp (ed.), *The Oxford Handbook of Ethical Theory*, Oxford: Oxford University Press.

Timmons, M. 1999. *Morality without Foundations*, Oxford: Oxford University Press.

Thagard, Paul. 1978. "The Best Explanation: Criteria for Theory Choice," *The Journal of Philosophy* 75(2): 76-92.

Trivers, Robert L. 1971. "The Evolution of Reciprocal Altruism," *The Quarterly Review of Biology* 46(1): 35-57.

van Fraassen, Bas C. 1980. *The Scientific Image*, New York: Oxford University Press.

Stich, Stephen, and Jonathan Weinberg. 2001. "Jackson's Empirical Assumptions," *Philosophy and Phenomenological Research* 62(3): 637-643.

Weston, Thomas. 1992. "Approximate Truth and Scientific Realism," *Philosophy of Science* 59(1): 53-74.

Westermarck, Edward. 1932/1960. *Ethical Relativity*, Paterson, NJ: Littlefield, Adams & Co.

Will, Clifford M. 1986. *Was Einstein Right?* New York: Basic Books.

Williams, Bernard. 1980. "Internal and External Reasons," reprinted in *Moral Luck*. Cambridge: Cambridge University Press (1981).

_____. 1985. *Ethics and The Limits of Philosophy*, Cambridge, MA: Harvard University Press

Zemach, E.M. 1976. "Putnam's Theory on the Reference of Substance Terms," *The Journal of Philosophy* 73(5), 116-127.