2013

A methodologically naturalist defense of ethical non-naturalism

Abraham Graber
University of Iowa

A METHODOLOGICALLY NATURALIST DEFENSE OF ETHICAL NON-NATURALISM

by

Abraham David Graber

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Philosophy
in the Graduate College of
The University of Iowa

August 2013

Thesis Supervisor: Professor Diane Jeske

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Abraham David Graber

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Philosophy at the August 2013 graduation.

Thesis Committee: ___________________________
Diane Jeske, Thesis Supervisor

____
Richard Fumerton

____
Evan Fales

____
Carrie Figdor

____
Ali Hasan

ACKNOWLEDGMENTS

I would like to thank Diane Jeske for all of her help and guidance over the years. I would also like to thank Richard Fumerton, Evan Fales, Ali Hasan, and Carrie Figdor for serving on my committee as well as for all of the support and feedback they have offered in the course of my tenure at Iowa.

This project would not have been possible without the friendship, support, and philosophical insight of a good many people. In particular, I owe a significant debt of gratitude to (in no particular order): Benjamin Berman, Andrew Parker, Kristopher Phillips, Samuel Taylor, Seth Jones, Daniel Schulz, Daniel Roston, and Ian O'loughlin. I entertain, with skepticism, the thought that I could have completed the dissertation without the help of each and every person on the list. I am particularly grateful to Jessica Schwartz for her support in the final stages of the project.

Last but not least, I would like to thank my family for their undying support. I would never have made it this far without the encouragement of Hetty Hall, Mark Graber, and Rachel Graber.

TABLE OF CONTENTS

CHAPTER ONE:
THE DIALECTICAL SPACE

Introduction

Within the philosophical community a certain kind of commonsensical view, whereby our moral judgments are about a special kind of robust truth, has largely fallen out of favor. Proponents of this commonsense view have come to be known, derisively, as "mad dog realists." I intend to introduce a new form of mad dog realism—even more ontologically rabid than its already unpopular kin—and argue that, if one is committed to a scientific worldview, one should also be a mad dog realist of this particularly rabid variety. These are lofty goals, the pursuit of which will require the entirety of this dissertation. For the time being, we would do well to aim at more modest targets.

Getting clear on the question being asked is key to success in any philosophical enterprise. Before I can offer a defense of mad dog realism, we must understand both the view and its primary opponents. Before we can do that, we must first get clear on what kind of question mad dog realism is supposed to be an answer to. These are my aims in the present chapter. By its conclusion I hope to have provided the reader with a rough sketch of the various kinds of questions that get asked as a part of ethical inquiry and a brief overview of the kinds of answers that have been given to the particular question I intend to address in this dissertation.

Normative ethics, applied ethics, and meta-ethics

There are, broadly construed, three fields of ethical inquiry. "Normative ethics" is, of the three, most likely to be familiar to the reader. Our lives are frequently filled with difficult moral choices. Some of these choices are relatively innocuous, e.g. is it morally permissible to tell your sister that you like her haircut, though in fact you do not? Other choices are more troubling: to what extent is it morally permissible to draw on the powers of my job to give a friend or relation special treatment? Casual reflection should reveal that confidence in your approach to familiar moral quandaries should not lead you to be confident that you can find the morally permissible path in novel situations. Specifics of a particular circumstance determine what actions are morally permissible. Morally relevant variables abound and minute changes in the configuration of these variables can have a profound impact on what one ought to do. Normative ethics is motivated by two central worries. First, how can I be confident that the individual moral judgments I make are accurate? Second, how can I be confident that the moral judgments I will make in novel circumstances are accurate? Given the importance we place on moral considerations, it would be quiet an understatement to describe having answers to these two questions as "nice."

The aim of normative ethics is to provide a response to each demand. A single account can serve as an answer to both questions. Both questions are underpinned by uncertainty regarding the nature of right action. Were we to have a principle for identifying when an action is right then application of this principle could both (1) give me confidence that the moral judgments I am currently committed to are correct and (2) in the future, give me guidance for how I ought to act. If one has the correct moral principle in hand, one need merely apply it to the circumstances one finds oneself in and, voilà, it will become clear what one ought to do. Normative ethics is constituted by the search for the principles that determine right action.

As I understand normative ethics, the paradigmatic normative ethical debate is the debate between the deontologist and the consequentalist. The consequentalist offers a principle along these lines: the right action is the action that, compared to alternatives, makes the world the best place. The deontologist takes a different approach. On one well-known view, the right action is the action that is in accordance with the predominant weight of one's *prima facie* duties (Ross). If the debate between the consequentalist and the deontologist can be settled, we will have come a great distance towards having a universally applicable guide for how we ought to behave.

Normative ethics is one of a triumvirate of areas of ethical inquiry. It is the area of ethical inquiry that the average person is likely most familiar with and it is likely the area of ethical inquiry that is most apropos to our daily lives. In an important sense, normative ethics lies between the other two areas of ethical inquiry: applied ethics and meta-ethics. The meta-ethicist asks questions about normative ethical claims whereas the applied ethicist relies on normative ethical claims to serve as premises in her arguments. It may be helpful, then, to consider both applied ethics and meta-ethics in light of the account of normative ethics I have offered. Given that meta-ethics is the focus of my dissertation, I will conclude my discussion of the areas of ethical inquiry there.

Applied ethics is, in an important sense, a step down in generality from normative ethics. The normative ethicist searches for the general principles that determine what we ought to do. If normative ethics is completed, presuming that we know all of the morally relevant facts, we will be able to determine in any given situation what we ought to do. The applied ethicist also seeks to give us guidance regarding how we ought to behave; however, the applied ethicist's project is not nearly as broad in scope. The applied ethicist looks to tell us what we ought to do in a constrained set of circumstances. I take medical ethics to be the most successful instance of an applied ethics research program. We would do well to look there for examples of the types of projects constitutive of applied ethics research.

One of the most contentious debates in contemporary medical ethics involves the permissibility of "conscientious objection." Proponents of conscientious objection argue that a physician is not obligated to provide any treatment the provision of which she believes is morally impermissible (Curlin). Opponents of the view disagree; the permissibility of refusing to provide treatment is independent of a physician's view about the moral status of providing said treatment (Card). Note the striking difference in scope between the questions asked by the normative ethicist and the applied ethicist. Whereas the normative ethicist searches for the universal principles that determine what we ought to do, the applied ethicist considers specific circumstances and attempts to determine, given these circumstances, what actions would be morally permissible. The primary difference between normative ethics and applied ethics is one of generality. Thus, it should come as little surprise that the distinction between normative ethics and applied ethics is not hard and fast. Instead, it is a question of degree.

By way of illustration, consider the reigning paradigm in medical ethics. The received view is that the physician must, in dealing with patients, balance the demands of four principles: beneficence, non-maleficence, autonomy, and justice (Childress and Beauchamps). The moral permissibility of a physician's action is determined by the balance it strikes between these four principles. A violation of any given principle can only be justified if the violation is necessary in order to respect one of the other principles. All of this talk of principles is evocative of normative ethics. Proposing general principles to guide action is certainly much closer to doing normative ethics than is asking about the moral permissibility of conscientious objection in medicine. Here the lines between normative and applied ethics begin to blur. In so far as principlism, the dominant paradigm in medical ethics, constitutes applied ethics, as opposed to normative ethics, it is because the four principles are taken to be specific to the physician-patient relationship. It may well be that the four principles offer helpful guidance outside of that context; however, once one makes this move, one is likely no longer doing applied ethics. We can, then, differentiate applied ethics from normative ethics in terms of the scope of the moral guidance that each seeks to offer. A successful normative ethics would offer moral guidance to any individual in any situation. A successful applied ethics would offer guidance to a restricted class of individuals in a restricted set of situations. The question of where normative ethics begins and applied ethics lets off is one of degree; there will be cases that cannot best be classified as either one or the other.

If applied ethics is less general than normative ethics, then meta-ethics is more general than normative ethics. Both normative and applied ethics asks, at one level of generality or another: "what ought we to do?" Meta-ethics asks questions about the projects of the normative and applied ethicists. Meta-ethical questions can helpfully be divided into three distinct subsections. First, meta-ethics asks *semantic questions*, i.e. questions about the meaning of normative language. Second, meta-ethics asks *ontological* questions. In offering answers to the semantic question, philosophers provide various sketches of what the world have to be like for it to be true (or false) that, e.g., *it is morally impermissible to cause severe pain for trivial enjoyment*. This is not yet enough to tell us if it is actually morally impermissible to cause severe pain for trivial enjoyment. We further need to know if the world meets the requisite conditions set out by some proposed semantic analysis. In asking the ontological question, the philosopher attempts to discover if the entities required for moral claims to be true, as specified by some semantic account, are part of our universe. Lastly, meta-ethics asks *epistemic* questions. Do we know that *it is morally impermissible to cause severe pain for trivial enjoyment*? If yes, how? If no, why not? These epistemic questions are important; however, of the three types of meta-ethical questions, I am least interested in the epistemic ones.

The meta-ethical terrain

This presentation of meta-ethics is highly schematic and likely largely unhelpful to a reader who is not already familiar with meta-ethical debates. My dissertation will focus on meta-ethics, in particular on questions of moral ontology. I will be arguing for ethical non-naturalism, known "colloquially" to philosophers as "mad dog realism." Non-naturalism is a member of a set of views that go by the name "moral realism." On my way to characterizing non-naturalism, I will first offer a characterization of "moral realism." It is easiest to understand the commitments of the moral realist against the backdrop of a general understanding of meta-ethical debates. In what follows I will provide a rough sketch of the meta-ethical territory. I will classify meta-ethical views by (1) their semantic analysis and (2) their ontological commitments. Semantic and ontological commitments are tightly connected. In many cases, the ontological commitments of a view can be read off of its semantic analysis. The semantic and ontological commitments of a meta-ethical view underdetermine the view's epistemological commitment. Thus, epistemological accounts can vary within a family of semantic/ontological views. The taxonomy of meta-ethical views is, like any taxonomy, a consequence of distinctions. I will start by presenting the most coarse-grained distinctions and progress to the more fine-grained.

Cognitivism vs. non-cognitivism

All meta-ethical views can be classified as either *cognitivist* or *non-cognitivist.* As a first take, the cognitivist holds that moral judgments, e.g. "infanticide is morally permissible," are either true or false. The cognitivist holds that moral claims are descriptions. The non-cognitivist disagrees. For the non-cognitivist, moral "judgments" are neither true nor false.

There are, broadly speaking, two kinds of non-cognitivists. One might be an *expressivist* or one might be a *prescriptivist.* Generally speaking, the expressivist understands moral judgments to primarily be in the business of expressing some affective state. The prescriptivist understands moral judgments to primarily be in the business of prescribing a course of action.

Stevenson's emotivism was an early precursor to contemporary expressivism. Stevenson thought that persuasion was the primary function of moral language. Consider a mother who says to her unkempt child: "We like taking showers." When said with the appropriate inflection and force, it is clear that the mother is not intending to convey the fact that *we like taking showers*. Instead, the utterance is intended to convince the child to shower—expressing approval of a certain kind of behavior as well as an expectation that this approval be shared. Stevenson takes moral language to be analogous. The primary meaning of standard moral utterances is nothing more than an expression of approval or disapproval; however, just as "we like taking showers" has a certain kind of persuasive force, so do moral utterances. On Stevenson's view, moral claims are no more true or false than is, e.g. "Boo infanticide!" However, unlike "Boo infanticide!" moral utterances come with a certain kind persuasive power, imbued on them by our practice of using moral utterances as a method for changing the behavior of others (Stevenson 18-19).

When compared to contemporary versions of expressivism, Stevenson's emotivism can seem crude. Gibbard offers a more nuanced expressivist account. For Gibbard, saying "infanticide is morally impermissible" expresses one's commitment to a norm of avoiding infanticide. Importantly, on Gibbard's account, "infanticide is morally impermissible" cannot be translated to, "I am committed to a norm of avoiding infanticide." The latter is a description and is thus either true or false. Instead, the utterance "infanticide is morally impermissible" should be understood as expressing commitment to a complicated set of behavioral dispositions that result in the avoidance or prevention of infanticide. Much like Stevenson, Gibbard thinks that expressing commitment to a norm can influence the behavior of those around you. Much as telling a friend that *you liked Casino Royale* may play an important role in convincing her that the film is *worthy of being watched*, expressing commitment to a norm can play an important role in convincing others that the norm is *worthy of being committed to*. Whereas Stevenson offers a relatively thin account of what it takes to sincerely make a moral utterance—approval of a certain type of action and the desire to make others approve of it as well—Gibbard offers a thicker account of the sincerity conditions for a moral utterance—acceptance of a complicated set of related behavioral dispositions (Gibbard).

Hare's prescriptivism contrasts with the expressivism favored by Stevenson and Gibbard. Hare takes moral judgments to be equivalent to the assertion of a prescription. Imagine two people protesting outside of an abortion clinic. One holds a sign that reads, "Abortion is Wrong!" Another holds a sign that reads "Stop Abortion!" According to the prescriptivist, these two signs should be taken to mean the same thing. Moral judgments are not descriptions, but prescriptions. When one makes a moral judgment, one prescribes a course of action (Hare).

Hare develops his prescriptivism along Kantian lines. For Hare, making a genuine moral judgment requires that the speaker be willing to universalize the relevant prescription. Suppose John genuinely asserts, "infanticide is wrong." Hare thinks that this utterance has the same meaning as the universal prescription, "don't engage in infanticide!" The prescription is universal in that it applies to everyone, in all places, at all times. For Hare, making a genuine moral judgment requires that one accepts a universal prescription. "Acceptance of a universal prescription" can be understood in terms of how one would behave under various circumstances. If there are conditions under which one would be willing to commit infanticide, one has not accepted the universal prescription *do not engage in infanticide!* One has only accepted this universal prescription if there are no circumstances in which one would kill an infant.

Whether expressivist or prescriptivst, the non-cognitivist's ontological commitments can be read off of their semantic account. The surface grammar of moral claims appears to be descriptive. If I say, "torturing innocents is morally impermissible," I appear to be predicating a property, i.e. *being morally impermissible*, to a subject, i.e. *torturing innocents*. The question then arises: what is the ontological status of the property *being morally impermissible*? No such question arises for the non-cognitivist. Asking about the ontological status of the property *being morally impermissible* belies, for the non-cognitivist, a significant conceptual confusion. Any (object level) claim about moral impermissibility is not a description. There is nothing in the world that corresponds to *being morally impermissible*, nor could there be. The non-cognitivist might, rightly, wonder what it could mean for there to be a property that corresponds to "Boo Abortion!" or

"Stop Abortion!" The ontological commitments of the non-cognitivist are a direct consequence of her semantic account: there is no such thing as a moral property.

Unlike the non-cognitivist, the cognitivist (also know as a descriptivist) thinks that moral utterances can be either true or false. That is, the cognitivist thinks that moral utterances *describe* the world. There is, however, significant variance amongst the views that qualify as versions of cognitivism.

Sentimentalism

S*entimentalism* constitutes a significant and popular family of descriptivist (or cognitivist) views. All sentimentalists share two commitments: (1) all sentimentalists are cognitivists and (2) all sentimentalists understand moral claims such that the truth of all moral claims are determined by the emotional states of some set of individuals.[1] Sentimentalism can be further subdivided into three categories. A sentimentalist might be a *radical subjectivist*, a *(cultural) relativist* or a proponent of an *ideal observer view*.

The radical subjectivist holds that moral claims are made either true or false by the affective reactions of a single agent. Radical subjectivists can be further divided into *appraiser subjectivists* and *agent subjectivists*. The appraiser subjectivist holds that moral claims are made either true or false by the sentiments of the speaker, e.g. the claim that "abortion is wrong" is true *for me*, just in case I have a sentiment of disapprobation towards abortion.[2] The agent subjectivist holds that moral claims are made either true or

[1] This claim is in need of refinement. As it stands, this characterization of sentimentalism will classify paradigmatically non-sentimentalist views as versions of sentimentalism. I will briefly discuss some of the reasons for the problem when I offer a characterization of *realism*. It may be best to classify sentimentalism via a disjunctive definition: one is a sentimentalist just in case one accepts one of the three sentimentalist views discussed below.

[2] Depending on the specifics of one's view, this may be either an occurrent sentiment or a dispositional sentiment.

false by the sentiments of the object of moral assessment, e.g. the claim that "Jimmy ought not have hit Jessie" is true just in case Jimmy has a sentiment of disapprobation towards hitting Jessie. Importantly, on the radical subjectivist view, the truth of moral judgments is relativized either to the appraiser or to the object of moral assessment. On the appraiser view, abortion can be wrong *for you*, if you feel a sentiment of disapprobation towards abortion, and permissible *for me*, if I do not. On the agent view, it may be true that "Jimmy ought not have hit Jessie" but false that "Rufus ought not have hit Jessie" for no other reason than that Jimmy feels disapprobation towards the hitting of Jessie and Rufus does not feel disapprobation towards the hitting of Jessie.

(Cultural) relativism is essentially a scaled up version of subjectivism. Whereas the subjectivist holds that moral statements are made true in virtue of the sentiments of a *single* individual, the relativist holds that the truth conditions of moral statements depend on the sentiments of more than one person. The relevant set of individuals is usually, but does not necessarily have to be, a culture. It is notoriously difficult to offer a plausible account of what is meant by "culture." There are a variety of related reasons that this is the case. Depending on how one defines "culture," intra-cultural differences in moral outlook are often significant. When there is no moral consensus in a culture, it is not clear what, on the cultural relativist's view, would make moral claims either true or false. One could attempt to solve the problem by defining cultures in terms of the moral views held by the members; however, doing so puts one in the difficult position of needing to find some criteria that will demarcate cultures based on the moral views of their members while not collapsing cultures into individuals, and thereby reducing relativism into subjectivism. Because of worries like these, relativism is not a popular position amongst moral philosophers.

Again, there are appraiser and agent versions of the view. An *appraiser relativist* might hold that the claim that "Jimmy ought not have hit Jessie," when said by S, is true if and only if there is a prevailing sentiment of disapprobation towards Jimmy's hitting of

Jessie amongst the members of S's culture. An *agent relativist* might hold that "Jimmy ought not have hit Jessie" is true if and only if there is a prevailing sentiment of disapprobation towards Jimmy's hitting of Jessie amongst the members of *Jimmy's* culture. Again, note that on the relativist view, the truth of moral claims is (unsurprisingly) relativized. An action can be *right* for Jimmy but *wrong* for Zackary, depending either on (1) the culture of the individual doing the judging or (2) the culture of Jimmy and Zackary respectively.

Ideal observer views are the final member of the sentimentalist family. A proponent of an ideal observer view holds that moral statements are made true by the counter-factual sentiments of a set of idealized agents. Firth offered the seminal presentation of the view. On his account, amongst a host of other conditions, idealized agents are required to be factually omniscient, possess maximally vivid imaginations, and be perfectly consistent. On an ideal observer account, the claim that "Jimmy ought not have hit Jessie" is made true by the sentiment of disapprobation an idealized observer would have towards Jimmy's hitting Jessie (Firth).

There is no clear distinction between idealized versions of subjectivism or relativism and ideal observer accounts. Many ideal observer views would, were they to include fewer counterfactual conditions, be best classified as subjectivist or relativist. Degrees of idealization are just that: degrees. There is logical space for any number of views that cannot easily be classified either as some version of subjectivism or relativism, *or* as an ideal observer view. A brief illustration might help to make the point clear.

Firth was the first contemporary philosopher to explicitly defend an ideal observer view. Firth's ideal observer is: "omniscient with respect to non-moral facts," omnipercipient, disinterested, dispassionate, consistent, and in all other respects, normal (Firth 115). On Firth's view, the agent whose sentiments make true moral judgments clearly deserved the label "ideal observer." Such an agent looks very little like the rest of us. Compare Firth's view to an account Jesse Prinz at one time defended:

> We should say that the word ‘wrong’ refers only to those things that irk me under conditions of full factual knowledge and reflection, and freedom from emotional biases that I myself would deem as unrelated to the matter at hand. (Prinz 35)

Notice the important similarities between the two analyses. Prinz does not go so far as to demand that the agent whose affective responses make moral judgments true be omniscient; however, Prinz does require that the agent have all relevant factual knowledge. Prinz does not require that the agent be either entirely dispassionate or disinterested; however, Prinz does require that the agent's affective responses not be, from the agent's perspective, morally irrelevant. Prinz's view is often taken to be an instance of subjectivism whereas Firth's view is universally considered to be an instance of an ideal observer account. The difference: a question of degree. Were we to add further counter-factual constraints to Prinz's agent, or remove some counter-factual constraints from Firth's agent, the two accounts would look very similar. In adding/removing various counter-factual conditions from a view, one will encounter no bright-line such that all and only ideal observer views fall on one side of it, and all and only versions of subjectivism and (cultural) relativism fall on the other side.

The ontological commitments of the sentimentalist follow immediately from the various sentimentalist semantic theses. The sentimentalist is a descriptivist. As such, the sentimentalist takes the surface grammar of moral utterances seriously. The claim that *murder is wrong* predicates a property, *being wrong*, of a subject, murder. The sentimentalist is committed to giving an account of the nature of moral properties. The moral properties are whatever the moral predicates refer to. For the subjectivist, moral predicates refer to the sentiments of individuals. For the (cultural) relativist, moral predicates refer to the prevailing sentiments within a culture. For the proponent of an ideal observer view, moral predicates refer to the sentiments of some hypothetical agent. Thus, for the proponent of any view that falls within the sentimentalist family, moral properties are nothing more mysterious than mental properties, often the sentiments of approbation and disapprobation. Looking for moral properties over and above these mental properties is rooted in a misunderstanding of the meaning of moral terms.

Constructivism

Sentimentalism is often taken to constitute a much too cavalier approach to morality. The worry is that, by attempting to analyze moral claims in terms of affective states, sentimentalist accounts threaten to degrade the seriousness of the moral endeavor. If sentimentalism is correct then there is a very tight analogy between, e.g., the fact that I dislike eggplant, and the fact that killing is wrong. Each fact is entirely constituted by non-rational affective responses. Reacting to this apparent threat to morality, philosophers have proposed competing semantic accounts that put significant stress on the importance of rational contemplation. *Constructivism* constitutes one of sentimentalism's most important competitors.

The constructivist offers an analysis of moral terms into folk psychological predicates associated, not with affective states, but with careful thought. A constructivist might think that *killing is wrong* just in case rational agents would *agree* or *choose* not to engage in killing or perhaps just in case rational agents would *endorse* or *consent to* a norm that forbids killing. Immanuel Kant has offered the best-known defense of constructivism. On at least one interpretation of Kant, telling a lie is morally impermissible because it cannot be rational to universally endorse the practice of lying. Though Kant is the most famous defender of constructivism, I will consider the work of another philosopher by way of illustrating the view. Kant exegesis is both too difficult and too contentious to be worth attempting here.

Instead, I will illustrate constructivism by considering a neo-Rawlsian meta-ethical account. I say "neo-Rawlsian" instead of "Rawlsian" because it is clear that Rawls did not take himself to be doing meta-ethics (Rawls 1980, 517-519). Nonetheless, if one were to treat Rawls' account as an analysis of "right" and "wrong," it would offer a paradigm instance of a constructivist account. On a neo-Rawlsian view, the permissibility of an action is determined by the action's conformity with a set of principles. An action is obligatory if it is demanded by a principle, permissible if neither demanded nor forbidden by a principle, and impermissible if otherwise. The important question becomes: what determines the content of the relevant principles? On the neo-Rawlsian view, the content of the relevant principles is determined by rational agreement amongst a set of agents. The agents are placed behind the veil of ignorance, i.e. they know nothing about their gender, race, ethnicity, socio-economic position, etc. As such, they should have no vested interest in, e.g., acting in ways that bolster patriarchal power structures. Each agent is at least as likely to end up on the wrong side of the patriarchy as s/he is to benefit from it. On such a picture, the principles that determine the moral facts are themselves determined by agreement amongst agents in the original position. In contrast with any version of sentimentalism, a neo-Rawlsian constructivist meta-ethics fixes the moral facts via processes that we tend to think of as paradigmatically rational.

Once again, we can read the ontological commitments of a view off of its semantic theses. The constructivist attempts to analyze moral predicates into folk psychological terms; in particular, cognitive or rational folk psychological states, in contrast to the conative or non-rational folk psychological states preferred by the sentimentalist. For the constructivist, moral predicates refer to folk psychological states. Moral properties are nothing more mysterious than, e.g., the state of having consented to a norm.

Realism defined

One might prefer constructivism to sentimentalism because the former, and not the latter, takes morality seriously. This same style of criticism can, itself, be leveled against constructivism. On the constructivist view, the moral facts are still *constructed.* Whether one is a non-cognitivist, a sentimentalist, or a constructivist, morality is, in some sense, a human artifact. The same cannot be said of views that count as versions of *moral realism*. *Moral realism* is notoriously difficult to classify. The moral realist holds that moral judgments are made true in roughly the same way that descriptions of the external world are made true. The claim that *mass bends space-time* is not true in virtue of any real or hypothetical person's mental states. The truth of the claim is entirely independent of anyone's existence. Had life never entered onto the scene in our universe, it still would have been the case that mass bent space-time. The realist thinks that moral truths are on par with this kind of scientific truth.

While this sort of metaphorical gesturing may be good enough to convey the nature of the realist's hypothesis, it will not do if our aim is to engage in careful philosophy. Cashing out the exact nature of the "robustness" of moral truths is no easy task. The analogy with physical facts breaks down in an important way. Suppose, as some philosophers do, that pain is the only thing with *intrinsic moral disvalue* and that pleasure is the only thing with *intrinsic moral value*. If this view is combined with some version of consequentalism, i.e. the view that one ought to do whatever maximizes the balance of intrinsic value over intrinsic disvalue, every moral claim is made either true or false by the mental states of some individual, i.e. by pain and pleasure. Realism is strikingly difficult to characterize because, on one hand, the realist wants moral truths to be robust in the same way that the claim that *mass bends space-time* is robust. On the other hand, the realist must allow for the possibility that various moral claims are made true by mental states.

I will offer two definitions of moral realism, criticize each, and finally offer my own definition. Shafer-Landau offers the following definition:

> There are moral truths that obtain independently of any preferred perspective, in the sense 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 (*Moral Realism: A Defense* 15).

Shafer-Landau's definition attempts to capture the perspective-independence of moral truths. The trouble is that he does so by ruling out the possibility that moral claims are made true by "the ratification [of the standards that fix the moral facts] from within any given actual or hypothetical perspective." Shafer-Landau's definition is too weak in that it lets in meta-ethical views that are clearly not realist.

Much depends on what Shafer-Landau has in mind by "ratification." I take the following to capture the standard usage of "ratify": The United States Congress ratified The Patriot Act. "Ratification," as it is standardly used, suggests explicit adoption. Think back to the earlier discussion of constructivism. Talk of "ratification" seems most at home in this kind of framework. It is natural to speak of the principles that ground moral truths in a neo-Rawlsian framework as having been *ratified*. If Shafer-Landau means to use ratification in this sense then his definition of realism will not rule out versions of sentimentalism. According to the subjectivist, my sentiments of approbation and disapprobation are the standards that fix the moral facts. I only "ratify" my sentiments of approbation and disapprobation in a stretched-to-near-breaking metaphorical sense.

The second definition of realism I will consider is offered by Brink. Brink holds that the moral realist is committed to two theses: "(1) there are moral facts or truths, and (2) these facts or truths are independent of the evidence for them" (17).

Providing a definition of realism is difficult because it is very plausible to think that *pain is intrinsically bad*. Why is punching Jimmy impermissible? Presumably, the answer is that punching Jimmy is impermissible because it causes Jimmy pain. It seems implausible for the realist to demand that the truth-maker of moral claims be independent of anyone's mental states; however, demanding that the truth-maker of moral claims be independent of anyone's mental states may seem like the only way to make good on the analogy between moral facts and facts such as *mass bends space-time*. Brink's definition may offer to provide a way to navigate between Charybdis and Scylla.

On one popular view, a person's evidence is composed entirely of things that are accessible to her (Conee and Feldman 54). Necessarily, a person's pain is accessible to her. It may, however, not be the case that, necessarily, the *badness* of a person's pain is accessible to her. This allows us to distinguish between the mental state "pain" and the moral property of *badness*. It is impermissible to punch Jimmy because punching Jimmy causes pain; however, this is not to say that the impermissibility of punching Jimmy is *identical* to facts about Jimmy's pain. The pain is accessible to Jimmy whereas the badness of the pain may not be. Brink may have successfully driven a wedge between the mental states that serve as the truth-makers for some moral judgments and the moral properties themselves.

Unfortunately, Brink's definition has problems of its own. Brink readily admits that MR does not adequately capture the sufficient conditions for realism. Imagine either an ideal observer view or something like a neo-Rawlsian constructivism. Truths about the sentiments of an ideal observer, or truths about what principles those behind the veil of ignorance would ratify, are independent of any evidence we might have. Our best evidence might suggest that, according to either view *such-and-such is morally impermissible*; however, we could just be wrong about the sentiments of an ideal observer or the principles that those behind the veil of ignorance would ratify.

Brink's definition can be rescued from this criticism. As stated, Brink's definition is ambiguous. Consider the second clause: "[the moral] facts or truths are independent of the evidence for them" (Brink 17). We could read Brink as saying that the moral facts are independent of *my* evidence for them, *our* evidence for them, or *anyone's* evidence for them. If we take the latter option and read "anyone" so as to include hypothetical agents, Brink can avoid the criticism I pushed in the previous paragraph. The sentiments of an ideal observer are accessible *to the ideal observer*. As such, Brink does not have to count the ideal observer view as a version of realism.

Disambiguating Brink's definition in this way will only serve as a stopgap. We need merely modify the original objection. Imagine a version of neo-Rawlsian constructivism whereby ratification happens by a blind vote. Principles are proposed, every agent behind the veil of ignorance writes down his/her vote on a piece of paper, then places the piece of paper in an urn. Receiving the most votes is constitutive of being a moral principle. Once all of the votes have been cast, some principles have been ratified. Nonetheless, the fact that these principles have been ratified is independent of anyone's evidence; no one has had the opportunity to count the votes. In many ways, this is a very odd version of constructivism. It seems very odd for a meta-ethical view to spend time stipulating the method of vote counting. Nonetheless, it seems clear that the view I have sketched is a version of constructivism, and *not* realism. Brink's definition will not do.

I am deeply skeptical that any successful definition of moral realism will be forthcoming. As I sketched the original challenge, a mix between *hedonism* and *consequentalism* appears to constitute a view whereby all moral judgments are made either true or false by the mental states of some set of agents. The trouble is that the realist needs some way to count the mix of hedonism and consequentalism in, while keeping out various versions of sentimentalism and constructivism. I think that this challenge is surmountable. One need only specify that moral judgments are not made true by any agent's attitudes *towards the object of moral assessment.* On both the sentimentalist and constructivist picture, moral truths are determined by an agent's or agents' attitudes *towards* some state-of-affairs or action. However, there is a related difficulty for providing an adequate definition of moral realism and it seems to me that this latter problem is insurmountable.

In the contemporary medical ethics literature there is currently a raging debate over *conscientious objection*. The proponent of conscientious objection thinks that it is morally permissible for a physician to refuse to offer any treatment that she thinks is morally forbidden to offer (Curlin). The opponent of conscientious objection thinks that the permissibility of a physician offering a treatment is independent of her attitude towards the permissibility of offering the treatment (Card). This debate in the medical ethics literature has a parallel at the level of normative ethics. Consider some agent A, and some action x, and some set of circumstances C. Suppose that A has no view about the moral permissibility of x in circumstances C. Further suppose that, in such a scenario, doing x in C is obligatory for A. We can now ask the following question: supposing A believes that doing x in C is morally forbidden, is it permissible for A not to do x in C? I tend to learn towards answering "no." An agent's view about the permissibility of some action does not influence the actual permissibility of so acting. Nonetheless, it is certainly epistemically possible that I am wrong about this. Just as a mix between hedonism and consequentalism appears to be a live position, so does the view that an agent's view about the moral permissibility of an action can influence whether or not the action is, for the agent, obligatory. Given that this is a live normative option, the realist needs to be able to accommodate it. The trouble is that, if the realist needs to accommodate this normative position, she cannot hold that moral truths are independent of an agent's attitudes towards the object of moral assessment.

There are, however, still moves the moral realist could make. So far I have taken the following approach to characterizing the ontological commitment of various meta-ethical views: first we offer an account of the truth-conditions of (object level) propositions containing moral predicates, then we read the ontological commitments of the view off of the semantic account. The problem we have so far faced is that the realist wants to allow that mental states play a role in making various moral propositions true. This complicates the task of reading the realist's ontological commitments off of her semantic account. Nothing commits us to this order of analysis. We could, alternatively, attempt to give an account of the nature of the properties the realist is committed to, thus avoiding the need to talk about truth, and then derive the realist's semantic analysis from her ontological commitments. Shafer-Landau, in offering a characterization of the realist's position in terms of truth, takes the semantic approach. Brink straddles the line by offering his definition in terms of "moral facts or truths."

Once we have abandoned the semantic approach, we are now free to attempt a direct characterization of the realist's moral properties. The following account of moral properties suggests itself: *realistically construed moral properties are not identical to any mental properties*. This nicely rules out versions of sentimentalism. The sentimentalist thinks that moral properties just *are* mental properties. The definition, however, fails to rule out constructivism. The constructivist can hold that moral properties are identical with some complicated set of properties regarding social institutions or relations. On such an account, moral properties are not identical with mental properties, though moral properties are *partially constituted* by mental properties.

It may seem like there is an easy solution. To rule out constructivism, let realistically construed moral properties be such that they are not identical to, nor partially constituted by, mental properties. I worry, however, that this condition is too strong. A realist could plausibly hold that the state-of-affairs of *it being wrong to hit Jessie* is partially constituted by Jessie's mental states. So long as they cashed out this partial constitution claim appropriately, we would not want to deny realist status to the meta-ethical theory. It does not appear that reversing the direction of analysis—first attempting to characterize moral properties and then reading the realist's semantic commitments off of her ontological commitments—gets us any closer to providing an adequate definition of moral realism.

For these reasons, I have recently become very pessimistic about the prospects of offering a viable definition of realism. Given that I intend to defend a version of moral realism, one might think that this sort of pessimism is devastating to my project. While I am not happy with my inability to offer a rigorous definition of moral realism, I do not find it overly worrisome. My primary aim is to provide an accurate description of a certain part of the world. I will be attempting to identify and describe a certain type of property. Questions about appropriate classification are secondary. The best I can do is provide a rough-and-ready description of moral realism. Lacking jointly necessary and sufficient conditions for realism, I will be unable to provide a knockdown case that the view I will develop constitutes a version of realism. I hope that the view I will defend is a paradigm instance of moral realism. If not, so much the worse for moral realism. Either way, we ought not let questions of categorization hinder our search for the truth.

In light of my pessimism regarding the prospects of offering an adequate definition of realism, I have chosen to offer a negative definition. That is, I will demarcate positions that count as version of realism by demanding that they *not* qualify as a version of anti-realism. In short, on the definition I am about to propose, a view is a version of realism just in case it cannot be correctly classified as one of the anti-realist views we have so far discussed.

My definition of realism will, unsurprisingly, follow the taxonomical pattern I've thus far established. The definition will consist of three distinct parts: two semantic theses and an ontological thesis. The realist is committed to all of the following:

1. Moral judgments are either true or false.
2. All versions of sentimentalism and constructivism are false.
3. Some logically atomic propositions that contain moral predicates in a non-opaque context are true.

The first two theses are semantic. The first thesis aims to rule out any non-cognitivist semantic account. The second thesis rules out cognitivist competitors to the realist's semantic thesis.

The third thesis is ontological. The final thesis picks out a class of propositions: logically atomic propositions that contain a moral predicate in a non-opaque context. The thesis then makes a claim about some members of this class of propositions: they are true. As such, (3) is a claim about the nature of the universe. The first and second theses tell us about the content of the propositions in question. The third thesis tells us that there is something in the universe that corresponds with the content of the propositions picked out in (1) and (2).

In the remainder of the dissertation, I will use the term "mind-independent" to pick out the kind of properties the semantic conditions established by (2) aim to identify. A mind-independent property is any property that is not constituted in the way that the sentimentalist or constructivist thinks that moral properties are constituted. A mind-dependent property is any property that is not mind-independent.

Ethical non-naturalism precisified

Mad dog realism, or less colloquially, ethical non-naturalism, constitutes a subset of the views that qualify as versions of moral realism. Ethical non-naturalism is characterized by a commitment to the existence of *sui generis* moral properties. Put another way, the ethical non-naturalist is committed to the existence of moral properties that are ontologically basic. It is not, however, entirely clear how to make sense of the non-naturalist's thesis. What is meant by "*sui generis*" or "ontologically basic?" The following looks like a very plausible strategy for getting clear on what is meant by "ethical non-naturalism:" figure out what is meant by "non-naturalism" more generally, then, in light of having a general definition of "non-naturalism," determine the meaning of the phrase "*ethical* non-naturalism." There are two places one might look for a clarification of the meaning of "non-natural." One might look to either the philosophy of science or to meta-ethics to find a distinction between naturalism and non-naturalism.

The first approach, whereby one looks for a general definition of non-naturalism, is unlikely to be successful. Stroud notes the following about the use of the word 'naturalism' in contemporary philosophy:

> The idea of "nature," or "natural" objects or relations… has been applied more widely, at more different times and places, and for more different purposes, than probably any other notion in the whole history of human thought… "Naturalism" seems to

> be … rather like "World Peace." Almost everyone swears allegiance to it, and is willing to march under its banner. But disputes can still break out about what it is appropriate or acceptable to do in the name of that slogan. And like world peace, once you start specifying concretely exactly what it involves and how to achieve it, it becomes increasingly difficult to reach and sustain a consistent and exclusive "naturalism." (Stroud 43)

In short, there is no received view amongst philosophers about how to draw the distinction between naturalism and non-naturalism. This need not worry us too much. It would be sufficient for our purposes if there is a received view amongst meta-ethicists regarding how to best understand the distinction. Unfortunately, things do not look much better on the meta-ethical side of things:

> There may be as much philosophical controversy about how to distinguish [ethical] naturalism from [ethical] non-naturalism as there is about which view is correct… Most often 'non-naturalism' denotes the metaphysical thesis that moral properties exist and are not identical with or reducible to any natural property or properties *in some interesting sense of 'natural'*. [emphasis added] (Ridge)

If one is looking for a consensus on how to best understand the distinction between naturalism and non-naturalism, the meta-ethics literature does not offer a promising starting place. The extent of the agreement in the meta-ethics literature is the entirely uninformative claim that the non-naturalist is committed to thinking that moral properties "are not identical or reducible to any natural property or properties *in some interesting sense of 'natural'*" [emphasis added] (ibid.). This leaves us in something of a bind. Neither meta-ethics nor the philosophy of science is in a position to offer a clear distinction between the natural and the non-natural.

The following strategy does not look promising: define "ethical non-naturalism" by decomposing the phrase into its parts, i.e. "ethical" and "non-natural," then determine the meaning of the phrase by determining the meaning of each component of the phrase. It does not appear that a definition of "non-natural" will be forthcoming. Instead of taking a decompositional approach, I will treat "ethical non-naturalism" as a single technical term. I have, thus far, following Shafer-Landau (and Moore), characterized ethical non-naturalism as the commitment to the existence of *sui generis* moral properties. My aim will be to provide a definition of "ethical non-naturalism," understood as a single technical term, then offer an account of the nature of non-natural properties that follows from the account of ethical non-naturalism I have provided. I think this is likely to be the most promising route for providing an adequate account of the ethical non-naturalist's commitments.

The simplest way to proceed is to examine the commitments of those views that are widely regarded as versions of ethical naturalism and ethical non-naturalism. This will give us some idea of what kind of views ought to be counted as versions of ethical non-naturalism, and what kinds of view ought to be considered versions of ethical naturalism.

Unfortunately, not even this exegetical approach appears to offer much promise in arriving at an adequate definition of ethical non-naturalism. In many ways, Russ Shafer-Landau is *the* contemporary defender of ethical non-naturalism. It would not be a stretch to claim that his *Moral Realism: A Defence* is responsible for reviving ethical non-naturalism as a defensible meta-ethical position. On the other side of the debate, David Brink is considered one of the key defenders of Cornell Realism—the paradigmatic formulation of reductive realism. Why is this a problem? As far as I can tell, Brink and Shafer-Landau have *exactly* the same view about the ontological status of moral facts. David Brink, the naturalist, holds that: "[M]oral facts and properties are constituted by, but not identical with, organized combinations of physical facts and properties" (Brink 179). Russ Shafer-Landau, the non-naturalist, thinks that: "[M]oral properties are always realized exclusively by descriptive ones. Just as facts about a pencil's qualities are fixed by facts about its material constitution… moral properties are always realized exclusively by description ones" (*Moral realism: a defence* 77).

This puts us in a difficult position. If we want to characterize ethical naturalism and ethical non-naturalism by examining the ontological commitments of naturalists and non-naturalists, we ought to start by examining those accounts that are taken to be paradigm instances of each. Shafer-Landau and Brink are not fringe authors. Their work constitutes paradigm instances of non-naturalism and naturalism, respectively. That they appear to share the same view about the ontological status of moral properties is enough to tank the project of relying on exegesis to arrive at a satisfactory distinction between the two views. No matter how one divides up the logical space, either Shafer-Landau will end up being a naturalist, or Brink will end up being a non-naturalist.

Given the apparent lack of any consensus regarding the distinction between naturalism and non-naturalism, I will offer a novel way to draw the distinction. However, we are still in some need of a method for helping us determine where the distinction ought to lie. It will not do to draw the distinction arbitrarily. I will characterize ethical non-naturalism by demanding that any view that counts as a version of ethical non-naturalism can make sense of one of the most well known and intuitively appealing objections to the view. The method I intend to employ is most easily illustrated via analogy.

In arguing for evidentialism—the thesis that epistemic justification strongly supervenes on evidence—Conee and Feldman have argued that non-evidentialist accounts of justification cannot account for the intuitive draw of skeptical arguments (Conee and Feldman, chapter twelve). The thought is that any adequate account of justification needs to be in a position to make sense of the fact that people have taken certain arguments to pose a genuine threat for the possibility of knowledge about the external world. If one's analysis of "justification" makes a solution to external world skepticism too easy, it seems likely that one is not talking about "justification," but has changed the subject.

This argumentative move on the part of Conee and Feldman is suggestive of a more general methodological point: an accurate analysis of a property or predicate must leave room for one to take seriously the problems traditionally associated with that property/predicate. If one's analysis of a property or predicate too easily dissolves related philosophical worries, it seems likely that one is not offering an analysis of the property/predicate one originally had in mind. This dictum will underlie my second method for attempting to draw the distinction between ethical naturalism and ethical non-naturalism. My account of ethical non-naturalism must offer the resources necessary to understand the intuitive appeal of objections to the view.

The argument from queerness has, historically, constituted one of the most significant challenges to ethical non-naturalism. Consider Mackie's seminal formulation of the objection:

> If there were objective values, then they would be entities or qualities or relations of a very strange sort, utterly different from anything else in the universe… [A] way of bringing out this queerness is to ask, about anything that is supposed to have some objective moral quality, how this is linked with its natural features. What is the connection between the natural fact that an action is a piece of deliberate cruelty—say, causing pain just for fun—and the moral fact that it is wrong? It cannot be an entailment, a logical or semantic necessity. Yet is not merely that the two features occur together. The wrongness must somehow be 'consequential' or 'supervenient'; it is wrong because it is a piece of deliberate cruelty. But just what *in the world* is signified by this 'because'? (Mackie 19-20)

Mackie suggests a handful of ways we might attempt to understand the precise nature of this "queerness." Non-naturalists have argued, successfully to my mind, that either the features Mackie points to are (1) not queer (e.g. supervenience) or (2) not a commitment of ethical non-naturalism. Nonetheless, the intuition remains. There is something queer about non-natural moral properties. Furthermore, one of the central motivations for ethical naturalism is that the ethical naturalist's moral properties are not ontologically queer. The charge of ontological queerness offers to provide a guide to drawing the distinction between ethical naturalism and ethical non-naturalism. Any view that falls on the naturalist side of the divide ought not have putative problems with queerness. Any view that falls on the non-naturalist side of the divide must have the resources to explain this worry from queerness.

As I understand ethical non-naturalism, the ethical non-naturalist is committed to thinking that moral facts are not identical with microphysical facts. This claim is, however, ambiguous. There are three different ways one can interpret talk about identities. Moral facts might be *type-type* identical with natural facts, moral facts might be merely *token-token* identical with natural facts, *or* moral facts might be what I will call *disjunctive-type* identical with natural facts.

The distinction between type-type identities and token-token identities is best illustrated via examination of positions that have been held in the philosophy of mind. J.J.C. Smart has defended a mind-brain identity theory. According to this view, *types* of mental states, e.g. pain, are identical to *types* of brain states, e.g. c-fibers firing. On such a view, "pain" and "c-fiber firing" pick out exactly the same *type* of thing.

J.J.C. Smart's mind-brain identity theory contrasts nicely with functionalism. The functionalist characterizes mental states functionally. On one sort of functionalist view, pain is identical to *whatever it is* that plays a certain kind of functional role, e.g. is caused by tissue damage, leads to avoidance behavior, etc. In humans, c-fiber firing might stand in the correct functional role whereas in some species of alien, f-fiber firing might stand in the correct functional role. Pain, understood as a kind, is identical to a certain functional role; however, since that functional role can be instantiated in any number of ways, pain is not identical with any *type* of brain state. Token instances of pain, however, *are* identical to token instances of brain states. Consider some instance of a human's pain, P_1, at a particular time, t_1. Suppose that in humans the functional role of pain is played by c-fiber firings. It follows that P_1 at t_1 was *identical* to a specific instance of c-fiber firing. That is, the token instance of pain, P_1, is identical to some token instance of c-fiber firing. The mind-brain identity theorist is committed to *type-type* identities between mental states and brain states. On such a view, the property of *being in pain* is identical to the property of *having one's c-fibers firing*. The functionalist rejects type-type identities between mental states and brain states. However, the functionalist thinks that mental states and brain states are *token-token* identical. Some token instance of pain is identical to some token instance of c-fibers firing.

There is a third type of identity claim one might make, uncomfortably positioned between type-type identity and token-token identity. I will call this kind of identity *disjunctive-type* identity. Again, the view is best clarified via example. For a long time it was believed that jade constituted a natural kind. This turns out not to be the case. There are in fact two minerals, jadeite and nephrite, that have similar macro-properties but instantiate these macro-properties in distinct microstructures. In light of this discovery one might be tempted to deny that jade is type-type identical with any set of natural facts; jade is neither identical to jadeite nor is it identical to nephrite. One might, instead, accept a token-token account whereby, though jade is not type-type identical to any natural kind, every instantiation of jade is identical to some token of jadeite or some token of nephrite. But suppose that one wants to hold onto type-type identity claims for jade. Is there a way to do so? The answer is (sort of), "yes." One might take the following to be a property: *being jadeite* ***or*** *being nephrite*. One can now preserve one's type-type identity. The property of *being jade* is identical to the property of *being jadeite* ***or*** *being nephrite*. Why did I say that disjunctive type identity is uncomfortably positioned between type-type identities and token-token identities? As the jade example hopefully illustrates, if one is willing to posit disjunctive properties, one can always get type-type identities from token-token identities. Consider a functionalist's account of pain. Some pain, P_1, is token identical to an instance of c-fibers firing, C-fiber$_1$. Some other pain, P_2, is token identical to an instance of H-fibers firing, H-fiber$_1$. To get oneself a type-type identity one need only collect all of the instances of pain into an enormous (likely infinite) disjunction of particular pain instantiations and hold that this disjunction is identical to the property of *being in pain*. Let PI_n pick out an event that is token identical to an instance of pain. One can get a type identity for pain as follows: Pain = $\{PI_1 \vee PI_2 \vee PI_3 \vee \ldots \vee PI_n\}$. If it seemed to you a bit too easy to arrive at a type-type identity, you are in good company. There are reasons to be skeptical that the disjunction on the right side of the identity constitutes a genuine property (Armstrong).

Which kind of identity claim is the ethical non-naturalist committed to rejecting? The rejection of type-type identities between moral facts and microphysical facts[3] is not sufficient to capture the intuitive sense that non-natural moral properties are queer. To see that this is the case, one need merely note that non-reductive physicalism in the philosophy of mind, largely in the form of functionalism, is not taken to be plagued by the problem of metaphysically queer properties. The non-reductive physicalist holds that mental facts are not type-type identical with microphysical facts, though mental facts are token-token identical with microphysical facts. Indeed, in order to avoid the charge of queerness, Shafer-Landau self-consciously develops his account of moral properties along the lines of non-reductive physicalism about the mental, holding that moral facts are token-token, but not type-type, identical with microphysical facts (*Moral realism: A defence* 77*)*. If one rejects type-type identities between moral facts and microphysical facts while accepting token-token identities between the same two sets of facts, moral properties ought to be no more "queer" than non-reductive mental properties. Given

[3] Given my earlier refusal to characterize ethical naturalism in terms of identity between moral properties and natural properties, my move here may seem somewhat odd. I chose not to characterize ethical naturalism in this way because of the lack of an adequate definition of "natural." Attempts to provide an adequate definition of "physical" and "microphysical" are plagued by similar problems. This incongruity deserves comment. If forced to attempt to define "microphysical," I would offer a disjunctive-type identity. Let "microphysical" pick out any member, or some set of members, of the disjunction of the theoretical entities and properties of contemporary atomic and sub-atomic theory. As science progresses, this definition of microphysical will fail to be adequate; however, unless there is a significant change in the direction of the development of our atomic and sub-atomic theory, the arguments I develop in later chapters ought to apply to any updates in the definition of "microphysical" needed to keep up with the changes in atomic and sub-atomic theory. Given the extent to which the arguments I develop in the later chapters rely on contemporary scientific theories, if there is a significant change in the direction of the development of our atomic and sub-atomic theories, it is likely that my central arguments will fail. Under such circumstances, it will no longer be important that I am not in a position to provide a better definition of "microphysical."

that non-reductive physicalism in the philosophy of mind is not taken to be committed to problematically queer entities, realist views like those espoused by Shafer-Landau and Brink should also not be taken to be committed to problematically queer entities. The ethical non-naturalist must be committed to some stronger rejection of identity between moral and microphysical facts.

As I will be using the term, the ethical non-naturalist is committed to the rejection of token-token identities between moral facts and microphysical facts. Rejection of token-token identities between moral facts and microphysical facts is sufficient to make sense of the intuitive appeal of the argument from queerness. If moral facts are not token-token identical with microphysical facts then moral facts cannot, in principle, be explained in terms of any set of microphysical properties. Every token instantiation of a moral property is, from the sub-atomic or atomic perspective, entirely novel. This leaves the non-naturalist with a genuine metaphysical puzzle. How can these novel properties emerge out of configurations of atomic and sub-atomic particles? The intuitive sense that moral properties are queer can be understood as a worry about this metaphysical puzzle.

My aim in this dissertation is to provide a defense of ethical non-naturalism. Ethical non-naturalism is characterized by the following four theses:

1. Moral judgments are either true or false.
2. All versions of sentimentalism and constructivism are false.
3. Some logically atomic propositions that contain moral predicates in a non-opaque context are true.
4. Moral facts are not token-token identical to microphysical facts.

Error theory

There is one final meta-ethical view that needs to be introduced. If ethical non-naturalism is the protagonist of this dissertation, then *error theory* or *nihilism* is the antagonist. The realist and the error theorist share semantic commitments. Just like the realist, the error theorist rejects non-cognitivist, sentimentalism, and constructivism. However, the error theorist rejects the realist's ontological thesis. The realist thinks that, in order for moral claims to be true, they must correspond with mind-independent properties. The error theorist agrees. The realist further thinks that at least some moral claims do so correspond. Here the error theorist disagrees. The error theorist thinks that there just are no robust moral properties with which moral claims correspond. While, in an important sense, close neighbors, in another sense, error theory and realism lie at opposite ends of the meta-ethical spectrum. Sentimentalists and non-cognitivists want to capture the importance of morality while backing away from the realist's ontological commitments. The error theorist does not bother with niceties: the realist is right about the meaning of moral terms and *there are no moral properties!* Every substantive (object level) moral claim is false. By the lights of error theory, the importance we give to moral considerations rests on a fundamental confusion about the nature of the universe. The realist wants it all, sentimentalists and non-cognitivists want some of it, and the error theorist is more than willing to do without.

Conclusion

This concludes the first chapter of my dissertation. I hope to have offered the reader both a broad strokes understanding of where my project fits into ethical inquiry broadly construed and have left the reader with some idea of what views are in competition with my preferred meta-ethical account. Furthermore, I hope to have offered a novel characterization of ethical non-naturalism. Any view that fits the exacting standards I have set for ethical non-naturalism deserves the label "mad dog realism." The demand that the ethical non-naturalist deny token-token identities between moral properties and microphysical properties is both novel and commits the ethical non-naturalist to the ontological queerness of non-natural moral properties. My primary aim in the dissertation is to defend the ontological commitments of mad dog realism—chapters three through seven are dedicated to this project. However, before I can defend the non-naturalist's ontological commitment, I owe the reader some defense of the non-naturalist's semantic account. I turn to this project in the next chapter.

CHAPTER TWO:
IN DEFENSE OF A REALIST SEMANTICS

Introduction

In the first chapter I offered a taxonomy of meta-ethical views and threw my lot in with realism. As I have defined the view, realism consists of three distinct theses. The first two are semantic:

1. Moral judgments are either true or false.
2. All versions of sentimentalism and constructivism are false.

The third is ontological:

3. Some logically atomic propositions that contain moral predicates in a non-opaque context are true.

I am most interested in defending the realist's ontological claim. My defense of the realist's semantic claim will be cursory, consisting of no more than this single chapter. This chapter can be broadly divided into two sections. In the first section I will offer considerations in favor of the realist's semantic understanding. In the latter half I will offer arguments against competing views. Many of the arguments I will offer have been presented before. My lack of focus on the semantic question requires comment. I have four reasons for glossing over the semantic debate.

First, there is only so much that can be accomplished in a single dissertation. Offering what is, to my knowledge, the first attempt to provide a scientifically plausible defense of the non-naturalist's ontological thesis is already a daunting task. In light of the enormity of this project, it is only reasonable to spend less time focusing on the semantic debate. Feel free to think of the dissertation, not as a defense of realism full stop, but a defense of the following conditional: if the ethical non-naturalist has the correct semantic analysis, then ethical non-naturalism is the correct meta-ethical position.

Second, there is an important sense in which all versions of anti-realism, other than error theory, attempt to find a middle ground between having it all and doing without. That is, realism offers a complete vindication of the importance of morality in our lives. We care about right and wrong, and good and bad because there are deep and robust facts about rightness and wrongness, goodness and badness. Alternatively, nihilism threatens to trivialize the role morality plays in our lives. Time spent worrying about doing the right thing is time wasted: everything is permissible. Many philosophers are wary of the realist's ontological commitments while not wanting to accept nihilism's bleak picture. Thus, various anti-realist semantics attempt to chart a middle course, free from realism's ontological commitments while still leaving some room for morality. But it is unseemly to attempt to have one's cake and eat it too. If realism is untenable then we should face up to the meaninglessness of our moral lives instead of attempting to find a stopgap.

Relatedly, I take it that one of the primary motivations for rejecting the realist's semantic theses is the fear that the realist's ontology is untenable. As the moral nihilist sees the world, worrying about questions like, "what is the good life?" and "what kind of person should I be?" represents a failure to face up to the reality of our universe. If one rejects the realist's ontological thesis, rejecting the realist's semantic theses is the only way one can make room for morality in our lives. In broad strokes, the strategy of my dissertation is to show that the realist's ontological commitment is compatible with a scientific worldview. I suspect that success in this project would largely dissolve resistance to the realist's semantic theses.

Finally, it is not clear to me that it is particularly important that the realist's semantic thesis be correct. As the reader will see by the conclusion of the fourth chapter, there are certain positions that would be correctly classified as versions of moral anti-realism that I am willing to endorse. If our focus on living morally is going to be vindicated, it does not need to be the case that there are properties that correspond with moral predicates. This is surely a surprisingly claim! A more in-depth defense of this claim will have to wait until the end of the fourth chapter; however, let me foreshadow some of what is to come. In brief, one can imagine that our moral predicates fail to correspond to anything in the universe—this is the error theoretic claim. It could, however, still be the case that, though our moral predicates fail to correspond to any properties, there are still properties that fit the realist's description of moral properties. Moral terms might fail to refer even though properties fitting the appropriate description constitute a robust part of our universe. So long as the right kinds of properties are out there, whether these properties are picked out by our moral predicates appears to be a mere terminological matter. Presuming that the right kind of properties are out there and further presuming that our moral language fails to refer to these properties, we are faced with the unpleasant task of engaging in a radically revisionary approach to moral language; we will have to take the necessary steps to ensure that moral predicates come to refer to the appropriate properties. Nonetheless, the core of the realist's view remains intact. Though there may be no *moral properties* there are still *schmoral properties* and, with a little bit of linguistic tinkering, these *schmoral properties* can do all of the work the realist originally wanted her moral properties to do.

In support of a realist semantics

Nonetheless, though I would be content accepting a position that falls short of moral realism, I would rather not be forced to do so. It is, therefore, important to provide the reader with some reason to accept the realist's semantic account. I intend nothing I say in this chapter to constitute a knockdown argument. My much more meager goal is to establish a presumption in favor of the realist's semantic theses. I take it that, given the strategy of my project, establishing such a presumption would be a significant accomplishment. If there is a presumption in favor of a realist semantics, and the primary motivation for resisting the realist's account of the meaning of moral predicates—the putative untenability of the realist's ontology—is removed, then the realist wins the semantic battle.

The first realist thesis is that sentences containing moral predicates are either true or false. This cognitivist thesis contrasts with the non-cognitivist's claim that sentences containing moral predicates are neither true nor false. Consider the following sentence: "Abortion is morally permissible." The surface grammar of the sentence is familiar. It appears to be the same as the surface grammar of "The car is red" or "Sammy is a mother." The sentence 'Sammy is a mother' has a subject and a predicate. If one were to say, "Sammy is a mother," one would be claiming that some subject, i.e. Sammy, has some property, i.e. being a mother. Note the similarity to "Abortion is morally permissible." On face, "Abortion is morally permissible" attributes some property, i.e. being morally permissible, to a subject, i.e. abortion. Sentences that attribute properties to subjects are descriptions. They are either true or false. It is either true or false that *Sammy is a mother*. Either Sammy is a mother, or she isn't! The same appears to be true about "Abortion is morally permissible." Either it is the case that abortion has the property of *being morally permissible*, in which case the sentence is true, or abortion has the property of *being morally impermissible*, in which case the sentence is false. This is not to say that non-cognitivist must be wrong. It is to say, however, that the non-cognitivist is swimming upstream. We certainly seem to be describing the world when we say, "Abortion is morally permissible." The non-cognitivist will have to marshal significant evidence to support her understanding of moral language.

The point can, perhaps, be put a bit more forcefully. I am convinced that when I say "It is morally impermissible to put a cigarette butt out in a baby's eye for trivial pleasure," I am saying something that is, minimally, truth evaluable. If you, like me, share this understanding of your own utterances then, before you accept her thesis, you should demand nearly incontrovertible evidence from the non-cognitivist. After all, it's possible, but not particularly plausible, that the non-cognitivist knows, better than you, what you mean.

It is more difficult to establish a presumption in favor of the realist's second semantic thesis. The realist holds that:

2. All versions of sentimentalism and constructivism are false.

I think that reflection on moral disagreement supports the realist's semantic theses. It is commonplace to argue that various aspects of disagreement support the realist's semantic account; realists tend to argue that various anti-realist semantics make genuine moral disagreement impossible. If one is a subjectivist then moral truths are indexed to individuals. Any apparent moral disagreement is just that: apparent. The truth conditions of the claim that "abortion is morally impermissible" are different for Sam than they are for me. On the subjectivist view, Sam can truthfully say, "abortion is morally impermissible" while I can truthfully say, "abortion is morally permissible." It looks like we are merely talking past one another. Similar problems seem to arise for other anti-realist meta-ethical accounts.

While there is a lot to be said for this style of objection to anti-realist semantics, it is not the criticism I intend to push. Instead, I will argue that the anti-realist has difficulty making sense of the felt importance of moral disagreement. By way of contrast, consider aesthetic disagreement.

There is a famous Pollack at the University of Iowa Museum of Art. Imagine two museumgoers examining the work. One finds it aesthetically pleasing. The other does not. Further suppose that this deviation in attitude becomes verbal; an argument breaks out. Is the Pollack beautiful? While, were I involved in such a dispute, I would be incapable of saying much in defense of my position, imagine that each of the museum goers is well educated in art criticism. Each can point to properties of the painting that support her thesis. One can even suppose that the argument gets heated. Nonetheless, it seems that it would be sufficient to end the debate were one of the parties to say, "I don't know what to tell you—I guess I just find it (not) beautiful, and you do not." At the end of the day, whether or not something counts as beautiful is a matter of taste. The above locution silences disagreement by pointing out the futility of argument. If the Pollack strikes me as beautiful and strikes Sam as ugly, there may be nothing more to say on the matter.

I think that reflection on aesthetic disagreement reveals a larger trend in our thought. If we think that the truth about some area of discourse is a matter of preference or convention, our fire for argument goes out. The example of the disagreement over the Pollack was intended to illustrate what I take to be a psychological fact about humans: upon recognition that truths about some domain are subjective or conventional, we lose interest in argument. In light of our recognition of the non-robustness of these kinds of facts, determining the fact of the matter ceases to seem important.

How does this relate to moral disagreement? Suppose that, instead of disagreeing about the beauty of a Jackson Pollack, Sam and I disagree about the moral permissibility of abortion. Suppose that I attempt to end the argument by saying, "I don't know what to tell you—I guess I just think abortion is morally permissible and you do not. That's all there is to it." Unlike in the case of aesthetic disagreement, I doubt that this would be enough to end the argument. Moral disagreements appear to be intractable in a way that disagreements about aesthetics and etiquette are not. Recognition of the non-robustness of facts in a domain appears to make us care less about those facts. Our passion for moral truth is largely unparalleled. As a consequence of moral disagreement, friendships are broken, societal movements are born, and wars are fought. The seriousness with which we treat moral disagreement is suggestive of our underlying attitude towards moral truth. Moral disagreement makes anti-realist semantics look implausible. In light of our relaxed attitudes towards non-robust facts, were the realist semantic theses incorrect, we would expect moral disagreement to be taken far less seriously.

Against anti-realist semantics

I hope to have provided some *prima facie* reasons to prefer the realist's semantic account. Admittedly, these reasons fall far short of offering a conclusive case. In what follows I will bolster the positive considerations I've introduced in favor of the realist's semantic theses with a series of arguments against various anti-realist positions.

Enoch against anti-realism

David Enoch has recently argued against various anti-realist semantics by showing that they have implausible normative consequences. Enoch's argument rests on the following normative thesis:

> IMPARTIALITY: In an interpersonal conflict, we should step back from our mere preferences, or feelings, or attitudes, or some such, and to the extent the conflict is due to those, an impartial, egalitarian solution is called for. Furthermore, each party to the conflict should acknowledge as much: Standing one's ground is, in such cases, morally wrong. (*Taking Morality Seriously* 19)

In support of IMPARTIALITY, Enoch asks us to consider the following case:

> We're spending the afternoon together. I want to go catch a movie I've been looking forward to seeing. You'd rather play tennis. But both of us really want to spend the afternoon together. How should we proceed? It seems clear that some loosely speaking egalitarian or impartial solution is called for. Perhaps we should flip a coin … [o]r perhaps we should take turns … But anyway, it would be wrong for me to stand my ground, and just insist that we go to the movie theater. Doing so – without some rather special further story, at least – would be wrong, unreasonable if anything is. (*Taking Morality Seriously* 17-18)

Why would my insistence on seeing a movie be morally reprehensible? Enoch thinks that, on any plausible normative theory, "[y]ou and I are, in a sense, equally morally important" (*Taking Morality Seriously* 18). Demanding that we go to the movie treats my preferences as if they are somehow privileged. Such disdain for the interests of others is morally reprehensible.

It is important to keep in mind how weak Enoch's claim is. Enoch is not claiming that, *ultima facie*, it is morally impermissible to demand that we watch the movie. Suppose that I am a consequentialist and I have good reasons to believe that, if we go to the movie, the world will end up being a better place than if we play tennis. Were Enoch's argument to rest on the claim that, under these circumstances, it would be morally reprehensible to demand that we go to the movie, his argument would have very little force. His point is not that, all things considered, demanding that we go to the movie would be morally reprehensible. Instead, Enoch thinks that there is a *prima facie* presumption that everyone's preferences count equally. While this presumption can be overcome, behaving as if there were no such presumption is morally questionable.

As an all things considered principle, IMPARTIALITY is obviously false. Consequently, the principle has relatively little intuitive draw. Significant reflection on Enoch's scenario is needed to make IMPARTIALITY intuitively plausible. Imagine that you were friends with someone who always demanded, in cases where preferences were in conflict, that you do what they wanted. I suspect that this friendship would not last long. Certainly, part of the dissolution of the friendship would be pragmatic. It is unclear why one would remain in a friendship that offered little by way of personal satisfaction. However, this does not, I think, capture the entire story. Sometimes the interests of friends grow apart. In such cases, friends grow distant but usually part on good terms. I can expect that the dissolution of a friendship with someone who refuses to give my preferences consideration will be significantly stormier. When asked why the friendship ended I am unlikely to respond, "we merely grew apart." Much more likely, I will say things like, "I discovered that she was not the kind of person I thought she was." My description of my ex-friend will likely use evaluative language: greedy, selfish, uncaring. I will feel that my ex-friend failed to give consideration to something she *ought* to have cared about. Notice that, here, normative language has entered the picture. I feel that my friend has behaved in a morally reprehensible way and, were someone to point out to me that I also fail to give others' preferences consideration, I would be ashamed. Though, when stated abstractly, IMPARTIALITY lacks intuitive pull, reflection on particular cases draws out the plausibility of the principle.

Suppose, then, that IMPARTIALITY is true. Enoch then asks us to consider, not mere disagreement in preference, e.g. between seeing a movie and playing tennis, but moral disagreement. Suppose that, as a good follower of Ayn Rand, Sam thinks that it is morally impermissible to do volunteer work. Sam thinks that doing whatever is best for him is morally obligatory. I disagree. Further suppose that Sam and I both want to spend time together; however, when trying to find a time that works for both us, I realize that there is an important volunteer opportunity that conflicts with the time that works best for Sam. Sam wants to spend our time together watching a movie; he thinks it would be morally impermissible to volunteer. I would prefer, on moral grounds, to spend this time volunteering. This case is importantly analogous to the movie vs. tennis case considered above. We have competing preferences for how we spend our time together. Unlike in the original case it seems, not just morally permissible, but even morally obligatory for me to stand my ground. If Sam refuses to join me in volunteer work, I ought to offer to reschedule with Sam and volunteer anyway. IMPARTIALITY holds with disagreements about mere preferences. It does not appear to be applicable regarding moral disagreement.

Enoch insightfully notes that, by the lights of many anti-realist semantics, moral disagreement *just is* mere disagreement in preference. Both non-cognitivists and subjectivists clearly hold this view. The relativist thinks that moral claims are made true by the sentiments of a prevailing majority of some subset of individuals. If my moral claims are made true by one subset of individuals and your moral claims are made true by a different subset of individuals, then putative moral disagreement is mere disagreement in preference. IMPARTIALITY rests on the plausible view that everyone's preferences ought, *prima facie*, to be given equal weight. Non-cognitivism, subjectivism, and relativism all hold, in one-way or another, that moral disagreements are mere disagreements in preference. As such, each view has the following untenable normative consequence: IMPARTIALITY holds of moral disagreement, i.e. in cases of moral disagreement it is morally impermissible to hold one's ground. We ought to reject non-cognitivism, subjectivism, and relativism because acceptance of such views commits one to untenable normative consequences.

Two anti-realist semantic accounts are not subject to Enoch's criticism. First, ideal observer views are immune. Moral disagreement is disagreement, not between your preference and mine, but about the attitudes of an idealized observer. Second, constructivism avoids the criticism. The constructivist holds that moral claims are a consequence, not of preference, but of rational agreement. On either view, moral disagreement cannot be reduced to mere disagreement in preference.

Against expressivism

Though Enoch's criticism gives us some reason to be suspicious of several of the leading anti-realist semantic accounts, I prefer to further bolster the case against the realist's semantic opponents. In the remainder of this chapter, I will offer two novel arguments. The first argument takes aim at the most popular version of non-cognitivism: expressivism. The second argument aims high, attempting to raise a general worry for sentimentalism and constructivism alike. By the end of this chapter I hope to have offered some reason to doubt each of the anti-realist semantics I introduced in the previous chapter. Notably, some anti-realist semantics receive less attention than others. In particular, I spend very little time arguing against prescriptivism. This lack of attention is justified by the relative unpopularity of the view. Given how few contemporary philosophers espouse prescriptivism, for dialectical reasons, this version of non-cognitivism deserves less attention than its anti-realist semantic competitors.

It would not be a stretch to say that expressivism has arguably become the dominant meta-ethical account. As Mark Schroeder puts it: "Blackburn and Gibbard's efforts [to revitalize expressivism] have gone so far toward making expressivism a respectable view that expressivism is now being widely applied across the 'core areas' of philosophy" (*Being for* 7). The popularity of expressivism makes it a prime target for evangelical moral realists such as myself. The embedding problem, also known as the Frege-Geach problem, challenges the non-cognitivist to make sense of the meaning of moral predicates embedded in logically complex sentences. Drawing on Mark Schroeder's work I argue that any solution to the embedding problem will require that the expressivist take moral utterances to express second-order mental states. Moral predicates can be multiply embedded within other moral predicates. This leaves the expressivist with the unpleasant choice of deciding between holding that humans are cognitive superbeings, capable of having eighth order mental states, or admitting that Geach was right and the expressivist cannot make sense of the meaning of moral predicates embedded in logically complex statements.

The expressivist is committed to two distinct theses:

1. Moral utterances *express* moral mental states.[4]

[4] One might reasonably wonder what I mean by "moral mental state." I intentionally leave this unspecified. The expressivist takes moral utterances to express some kind of mental state. What kind of mental state do they express? The moral ones. Which mental states are the moral ones? The expressivist's answer to this question will vary depending on the specifics of her view. An expressivist might take any of the following first order mental states to constitute moral thoughts: approval, disapproval, guilt, shame, approbation, disapprobation, etc. Furthermore, an expressivist might take moral thoughts to be constituted by any combination of these first order mental states into second order mental states, e.g. approval of guilt, disapproval of approbation, etc. No easy precisification of "moral mental state" is available—everything depends on the specifics of one's preferred expressivism.

2. Moral thoughts are non-cognitive, i.e. a moral thought does not represent states-of-affairs in such a way that, necessarily, the mental state is deficient if it fails to accurately portray the world.

In contrast with earlier versions of non-cognitivism, expressivism is not primarily a thesis about moral semantics. Early non-cognitivists took themselves to be offering a meaning analysis of moral sentences (Ayer; Stevenson; Hare). The early non-cognitivists' account of moral thought followed from their account of moral semantics. Expressivism turns the order of analysis on its head. The expressivist is primarily interested in offering an account of moral thought. Her understanding of moral semantics is secondary, falling out of (1), (2), and an account of the nature of expression (Schroeder *Being for*). Expressivism is the primary target of this section. Other versions of non-cognitivism may escape my critique, depending on whether their solution to the negation problem forces them to take second-order mental states to underlie moral utterances.

Non-cognitivists of all types are faced with the embedding problem. The initial worry is easy to understand. Take a naïve version of expressivism whereby the utterance "killing is wrong" expresses disapproval of killing. Consider the following argument:

1. If killing is wrong, then killing Sam is wrong.
2. Killing is wrong.
3. Therefore, killing Sam is wrong.

As Geach notes, for this argument to be valid, "killing is wrong" must mean the same thing in each premise (Schroeder "What is the Frege-Geach Problem"). It is difficult to see how the expressivist can get this result. If "killing is wrong" means "Boo killing!" it is not clear how to make sense of the first premise of the above argument: "if boo killing, then boo killing Sam?" The embedding problem leaves the expressivist in need of some account of how unembedded moral predicates could mean the same thing as moral predicates embedded in logically complex sentences.

The embedding problem can take at least two different forms corresponding to two different ways in which sentences can be logically complex. The expressivist owes some account of the meaning of moral predicates embedded in higher-order propositions, e.g. "Sam believes that *killing is wrong*." The expressivist further owes some account of the meaning of moral predicates embedded in propositions containing first order logical connectives, e.g. "killing is wrong *and* stealing is wrong." I will focus on this second kind of logical complexity.

The expressivist has a ready first response to the embedding problem: "[N]ormative sentences have the same meaning when embedded as when unembedded because the meaning of the complex sentence is a *function* of the meaning of its parts" (Schroeder *Being for* 20). This is an exact analog of the move used to make sense of the meaning of embedded descriptive terms (Hare). Thus far, the expressivist has only offered a schema for solving the embedding problem. Details are still owed. The embedding problem is not univocal; the expressivist must offer a solution for each of the logical connectives.

The expressivist has *prima facie* problems capturing negation. The problem is that "murder is wrong" and "murder is not wrong" are incompatible and the expressivist cannot capture this incompatibility in terms of relations between propositions. The expressivist account of incompatibility will have to draw on properties of mental states. There are two strategies one might take in attempting to show that two mental states are incompatible. Two mental states might be incompatible as a result of entertaining incompatible content, e.g. *intending p* and *intending ~p*. Alternatively, two mental states might be incompatible attitudes towards *the same* content, e.g. it is incompatible to both *approve of murdering* and *disapprove of murdering*.

One might take the first strategy and hold, plausibly, that "disapproving of p" and "disapproving of not p" are incompatible. Suppose Jon says "Murdering is wrong" and I have good reason to believe that he meant it. This gives me good reason to believe that *Jon has the mental state expressed by "murdering is wrong."* Consider all of the scopes for the negation:

(y) Jon has the type of mental state expressed by "murdering is wrong."

(m1) It is **not** the case that Jon has the type of mental state expressed by "murdering is wrong."

(m2) Jon does **not** have the type of mental state expressed by "murdering is wrong."

(m3) Jon has the type of mental state expressed by "**not** murdering is wrong."

(m4) Jon has the type of mental state expressed by "murdering is **not** wrong."

Note that (m1) and (m2) are equivalent—both do not need to be included. For ease of expression, I will use "thinks that…" as shorthand for "has the type of mental state expressed by…" Making these changes nets us the following four propositions:

> (w) Jon thinks that murdering is wrong.
> (n1) Jon does **not** think that murdering is wrong.
> (n2) Jon thinks that murdering is **not** wrong.
> (n3) Jon thinks that **not** murdering is wrong.(Schroeder *Being for* 45)

The expressivist needs to replace "thinks that" with some other mental state ascription—by current hypothesis, disapproval. For any such replacement to be successful it must be the case that the meaning of (n1)-(n3) can be captured if one replaces "thinks that…" with the mental state putatively expressed by "murdering is wrong." Note that, in (n1)-(n3), there are three distinct places to insert the "not." By contrast, consider the following "translation":

> (w*) Jon disapproves of murdering.
> (n1*) Jon does not disapprove of murdering.
> (n2*) ???
> (n3*) Jon disapproves of not murdering.(Schroeder *Being for* 45)

There are three possible places to add a negation to "Jon thinks that murdering is wrong." There are only two places to add a negation to "Jon disapproves of murdering." It follows that the meaning of "Jon has the mental state expressed by murdering is wrong" cannot be captured by "Jon disapproves of murdering." Call this *the scope problem*. Note that the problem is not unique to "disapproves." It will arise for any non-descriptive first-order mental state.[5]

One might attempt to solve the scope problem by taking the second strategy. *Approval of x* and *disapproval of x* are incompatible mental states. If "x is morally obligatory" expresses approval of x, and "x is morally impermissible" expresses disapproval of x, then one is in a position to explain the incompatibility of "x is morally obligatory" and "x is morally impermissible." This strategy, however, fails to solve the scope problem. Note that "approval of not murdering" and "disapproval of murdering" fail to capture exactly the same scope:

(w*) Jon disapproves of murdering

(n1*) Jon does not disapprove of murdering.

(n2*) ???

(n3*) Jon disapproves of not murdering.(Schroeder *Being for* 45)

(w#) Jon approves of not murdering.

(n1#) Jon does not approve of not murdering.

(n2#) ???

[5] This is a reconstruction of Schroeder's re-construction of Unwin's criticisms of Blackburn and Gibbard (*Being for*). Unwin targets his objection at specific versions of expressivism ("Quasi-realism, negation and the Frege-Geach problem," "Norms and negation: a problem for Gibbard's logic"). Schroeder clearly thinks that the problem applies to every expressivist; however, his reasons for thinking so are a bit opaque. I hope to have, here, shown why this formulation of the negation problem must be dealt with by every expressivist.

(n3$^{\#}$) Jon approves of murdering.

The trouble is, in Schroeder's words, that "we need at least one attitude from the {'permissible', 'unobligatory'} pair *and* at least one from the {'impermissible', 'obligatory'} pair, both of which we take as primitive, in order to define sentential negation" (Schroeder *Being for* 47).

The failure of the above attempt to deploy the second strategy should not deter us from trying to find some other instantiation of the strategy that will be successful. If the combination of approval and disapproval fails to capture the required scope of the negation, perhaps some other combination of mental states can. Following Schroeder, let's introduce a new mental state. Call it "tolerance"(Schroeder *Being for*). We can posit that tolerance is such that it captures the missing scope of the negation:

(w$^{\%}$) Jon disapproves of murdering

(n1$^{\%}$) Jon does not disapprove of murdering.

(n2$^{\%}$) Jon tolerates murdering.

(n3$^{\%}$) Jon disapproves of not murdering.

But as Schroeder points out, this leaves the following question unanswered: why is Jon's disapproving of murdering incompatible with his tolerating murdering? It is clear that *approving of x* and *disapproving of x* are incompatible. It is not clear that *disapproving of x* and *tolerating x* are incompatible. The point of the negation problem is that the expressivist has difficulty building the logical relation of negation out of the mental states that she has available to her. The introduction of tolerance does not demonstrate that the expressivist can be successful in answering this challenge. Instead, it is, by fiat, to declare success (Schroeder *Being for*).

In merely demanding that the expressivist offer an account of the incompatibility of *disapproval of x* and *tolerance of x*, I suspect that Schroeder is letting the expressivist off too easily. "Permissible," "impermissible," "obligatory," and "unobligatory" can all be inter-defined. Positing "tolerance" has not increased the scope of "disapproval of murdering." Instead of adding an extra scope marker, we've cheated by adding a new predicate to account for the missing spot for the negation. If "tolerance" is supposed to pick out any member of the set [permissible, impermissible, obligatory, unobligatory] then all and only those moral propositions that can be expressed with "disapproval" ought to be expressible with "tolerance" and some configuration of negations. But if all and only those propositions that can be expressed with "disapproval" are expressible with "tolerance" and some configuration of negations, then tolerance cannot fill in the scope that "disapproval" failed to capture. Either "tolerance" fails to solve the scope problem *or* "tolerance" and "disapproval" cannot capture all and only the same moral propositions and the expressivist is committed to thinking that permissibility, impermissibility, obligatory, and unobligatory cannot be inter-defined.

So far, the prospects for expressivism look grim. No attempted instantiation of either strategy has found any traction against the negation problem. This is not, however, to say that no solution can be found. Thus far, all attempted solutions have taken the utterance "murdering is wrong" to express first-order mental states. It is very doubtful that any solution to the negation problem, in terms of first-order mental states, will be forthcoming. If, however, one is willing to take moral utterances to express second-order mental states, the negation problem can be solved. Take the first strategy: "x is morally impermissible" and "x is morally permissible" are incompatible because the two utterances express mental states that are incompatible in virtue of the incompatibility of the *content* of the two mental states. Following Schroeder, suppose there is some second-order mental state that, for the time being, we will label "being for" (*Being for*). "Being for" stands in for a mental state to be precisified by the expressivist at a later date.[6] Consider the following expressivist translation:

(w$^{@}$) Jon is for disapproving of murdering.

(n1$^{@}$) Jon is not for disapproving of murdering.

(n2$^{@}$) Jon is for not disapproving of murdering.

(n3$^{@}$) Jon is for disapproving of not murdering.

If the expressivist takes the first strategy and treats utterances that include moral predicates to express second-order mental states, the negation problem can be solved.

It is worth noting that, from a purely syntactic standpoint, the expressivist does not need to take moral utterances to express second order mental states. One can replace "disapproving" in (n1$^{@}$)-(n3$^{@}$) with any verb and end up with the required number of scopes for the negation, e.g.:

(w$^{\sim}$) Jon is for avoiding murdering.

(n1$^{\sim}$) Jon is not for avoiding murdering.

(n2$^{\sim}$) Jon is for not avoiding murdering.

(n3$^{\sim}$) Jon is for avoiding not murdering.

While one can insert any verb to solve the purely syntactic problem, there is reason to believe that only folk psychological verbs will allow the expressivist to have a plausible moral semantics. The crux of the problem is that we want moral judgments to be orthogonal to our decisions about how to act. To see that this is the case, consider the paradox of hedonism. The egoistic hedonist thinks that the right action is the action that maximizes the actor's happiness. However, if one always decides to act so as to maximize one's happiness, one may, as a consequence, live a less happy life. It is

[6] It may help, in understanding the nature of the expressivist's solution to the negation problem, to think of "being for" as "approval." "Jon is for disapproving of murdering" can be read as "Jon approves of disapproving of murdering."

plausible to think that having certain kinds of close relationships is essential to living the happiest life possible. It is also plausible to think that these kind of relationships cannot form if one is always deciding how to act based on what would bring oneself the most happiness. This puts the egoistic hedonist in the position of thinking that, though it is obligatory to act only in ways that maximize one's own happiness, one should nonetheless aim at being the kind of person who sometimes fails to act to maximize one's own happiness. Thus, the egoistic hedonistic might *be for pursuing other people's happiness* even though, on her view, this may be morally impermissible. While I am not an egoistic hedonist, I nonetheless think that the position is coherent. If the expressivist solves the scope problem by introducing a verb that does not correspond with some folk-psychological state—e.g. "avoiding" as in (n1˜)-(n3˜)—the expressivist is committed to thinking that every sincere moral judgment constitutes the endorsement of some decision regarding how to act. Consequently, the expressivist is committed to thinking that the position of the egoistic hedonist (or some analog position) is conceptually incoherent. This is an untenable result.

I would like the objection I offer in the next section to constitute the coup de grâce culminating from Schroeder's more than capable drubbing of expressivist semantics. The objection is only devastating to expressivism as a family of views if a solution to the negation problem requires that expressivists take moral utterances to express second order mental states. I have just argued that every plausible solution does; however, even if I am wrong, my argument still has important consequences for the contemporary debate. The dominant expressivist accounts take moral utterances to express second-order mental states. Thus, Blackburn (*Ruling Passions*, "Anti-realist expressivism and quasi-realism") has defended a view whereby "to think that x is wrong is to disapprove of x and to disapprove of those who fail to share this disapproval" (Sinclair 137). In Gibbard's influential *Wise Choices, Apt Feelings*, he argues that sincere moral utterances express acceptance of a norm. Furthermore, norms are (at least) partially constituted by mental states. On the view Gibbard endorses in 1990, moral utterances express second order mental states (*Wise choices, apt feelings*). The case is somewhat more convoluted with regard to the position Gibbard defends in his *Thinking How to Live*. There Gibbard appears to prefer an analysis whereby claims about rationality express first order mental states (*Thinking how to live*).[7] Nonetheless, his analysis of moral utterances still commits him to thinking that they express second-order mental states: "to think that compassion is good is to *accept a norm* that says to desire compassion" (*Thinking how to live* 7). At its strongest, my worry poses a problem for all versions of expressivism. At its weakest, my worry poses a problem for those expressivist views that are currently in vogue.

Much of the debate over the plausibility of expressivism has revolved around the expressivist's ability to solve the embedding problem. I will take a different tack. Instead of arguing that an expressivist semantics is a non-starter, I will argue that any expressivist solution to the embedding problem will commit the expressivist to implausible *psychological* consequences.

Some physicians have claimed a right to conscientious objection, i.e. some physicians have claimed that it is permissible for them to refuse to offer treatments that they find morally objectionable (Curlin). Imagine that Ian thinks that abortion is morally impermissible. Further suppose that Ian is engaged in a debate with Sam who thinks that physicians ought to be legally bound to offer abortions. From Ian's perspective, Sam's attitude seems to be that *it is permissible to force someone to do the wrong thing*. It seems likely that Ian would find this attitude of Sam's not just objectionable, but morally so. In virtue of finding Sam's attitude morally noxious, it seems likely that Ian will think that it is a

[7] Though in *Thinking How to Live* Gibbard's account of the mental state expressed by the rational "ought" may be terms of first order mental states, it also seems to immediately fall prey to the negation problem sketched above.

bad thing that Sam has this attitude. So it seems likely that Ian will believe the following: *it is bad that Sam thinks it is permissible to force someone to do the wrong thing.* A descriptivist will take this to be a second order mental state: {Ian thinks it is bad that[Sam thinks it is permissible to *force someone to do the wrong thing.*]} But remember that, for the expressivist, moral predicates express second-order mental states. The mental state in question contains three moral predicates: "bad," "permissible," and "wrong." Thus, what the descriptivist takes to be a second-order mental state the expressivist must think is sixth-order. The expressivist translates the mental state expressed by "Ian believes that *it is bad that Sam thinks it is permissible to force someone to do the wrong thing"* as follows:

> {Ian being-for[disapproval of[Sam being-for[not disapproval of[forcing someone to do ∃x(Ian being-for[disapproval of x])]]]]}

For the expressivist, if one genuinely asserts a sentence involving a triply embedded moral predicate, one must have a sixth order mental state.

I am not clever enough to come up with an instance of quadruply embedded moral predicates; however, one can further increase the order of the mental state expressivists are committed to by placing a triply embedded moral predicate inside further folk psychological ascriptions. Again, suppose that Ian and Sam had a debate over the permissibility of forcing physicians to perform abortions. Ian tells us that he believes that *it is bad that Sam thinks it is permissible to force someone to do the wrong thing.* Further, suppose Ian is something of a passive-aggressive fellow. Whenever Sam is just within range of hearing, Ian steers his current conversation around to the subject of the impermissibility of thinking it is permissible to force someone to do the wrong thing. In light of his behavior, it seems reasonable to form the following belief: "Ian intends for Sam to believe that Ian thinks that *it is bad that Sam thinks it is permissible to force someone to do the wrong thing.*" This appears to be a belief that I am capable of holding, though just barely. The moral descriptivist is committed to thinking that the mental state expressed by the above is fourth-order:

> {Ian intends for Sam to [believe that Ian [thinks that it is bad that Sam [thinks it is permissible to force someone to do the wrong thing.]]]]}

The expressivist is committed to thinking that the relevant mental state is *eighth order*:

> {Ian intends for Sam to [believe that [Ian being-for[disapproval of [Sam being-for [not disapproval of [forcing someone to do ∃x(Ian being-for[disapproval of x])]]]]]]}

It seems clear to me that only a cognitive superbeing would be capable of entertaining the above listed *eighth order* mental state. While I am convinced that I can believe that "Ian intends for Sam to believe that Ian thinks *that it is bad that Sam thinks it is permissible to force someone to do the wrong thing*," I am also convinced that the above listed monstrosity of an eighth order mental state fails to correspond to any mental state I am capable of having. If I am correct, then expressivism must be incorrect; it entails the false conclusion that I am a cognitive superbeing.

It likely behooves me to say more. The reader may not find introspective evidence regarding mental states as convincing as I. For such a reader, does my argument present good reason to be doubtful of expressivism? I think that the answer is still "yes."

I have argued that the expressivist is committed to the existence of eighth order mental states. The dedicated expressivist ought ask herself this question: prior to reading this paper, what was her stance regarding our capacity for entertaining eighth order mental states? I presume that, like the rest of us, the expressivist had no reason to suspect we were capable of having eighth order mental states. The expressivist can escape the force of the present criticism by simply adding the capability of having eighth order mental states to her account of human psychology; however, such a move looks quintessentially *ad hoc*. One can always dodge criticism by adding epicycles to one's theory. This is not to say that doing so is a good way to proceed.

More importantly, *ad hoc* additions to ontology make one's theory less super-empirically virtuous. It is widely accepted that, *ceteris paribus*, simpler theories are better theories. An account of moral semantics that does not commit us to the existence of eighth order mental states is, *ceteris paribus*, better than a theory that does. There is a broad range of analyses of moral semantics with less demanding psychological commitments. These psychologically simpler competitors give us good reason to be hesitant about expressivism.

While both of these points are worthy of consideration, I am doubtful that the committed expressivist will find either particularly worrisome. Real trouble arises when we consider the relationship between philosophy—in particular an account of moral semantics—and psychology or cognitive science. I take it that a commitment to naturalism is one of the primary motivations for expressivism. Non-cognitivism promises to net one a purely naturalistic account of morality without forcing one to commit the naturalistic fallacy. Part of being a naturalist is thinking that science, not philosophy, offers the best guide to discovering the entities with which we share this universe. This commitment to science as the best guide to truth holds equally well with regard to claims about human psychology as it does with regard to claims about the fundamental constituents of the physical world. It is, presumably, in light of some such commitment that the expressivist feels comfortable being committed to the existence of eighth order mental states despite the fact that, if there are any such states, they appear to be introspectively inaccessible. Though I know of nothing about expressivism that logically commits the expressivist to the following methodological claim, I am convinced that, as a matter of fact, expressivists tend to hold that: if, for some domain, x, there is some science, S, that studies x, then S, not philosophy, ought to dictate our beliefs about what entities exist in x. The expressivist can avoid the force of my original criticism by positing the existence of eighth order mental states. But this is to let philosophy, and not psychology or cognitive science, hold the reigns. In short, the following looks to be an empirical question: are we capable of having eighth order mental states? In positing the existence of eighth order mental states in order to save her view, the expressivist appears to be doing either *a priori* psychology or *a priori* cognitive science. If one is a committed naturalist, one ought to be doubtful of either kind of project. The best the expressivist can do is declare agnosticism about moral semantics and wait until the empirical verdict is in.

There is some reason to think that the empirical verdict is already in and, as things look right now, the verdict is not a happy one for the expressivist. We have some evidence regarding the highest order mental state humans are capable of having. It is likely difficult to imagine how someone might go about gathering empirical data relevant to this question. The trick is to ask questions about vignettes such that, in order to answer the question correctly, one must have an n^{th} order mental state. By way of illustration, consider the following vignette. The vignette and subsequent questions about the vignette were presented to subjects:

> Emma worked in a greengrocer's store. She wanted to persuade her boss to give her an increase in wages. So she asked her friend Jenny, who was still at school, what she should say to the boss. "Tell him that the chemist near where you live wants you to work in his shop." Jenny suggested. "The boss won't want to lose you, so he will give you more money" she said. So when Emma went to see her boss that is what she told him. Her boss thought that Emma might be telling a lie, so he said he would think about it. Later, he went to the chemist's shop near Emma's house and asked the chemist whether he had offered a job to Emma. The chemist said he hadn't offered Emma a job. The next day the boss told Emma that he wouldn't give her an increase in wages, and she could take the job at the chemist's instead…
>
> (a) Jenny thought the boss would believe Emma's story
>
> (b) Jenny knew the boss would not believe Emma's story…
>
> (a) Emma thought the boss believed that the chemist wanted her to work for him
>
> (b) Emma thought the boss knew that the chemist had not offered her a job…

> (a) Jenny thought that Emma believed that the boss knew that the chemist did not want Emma
>
> (b) Jenny thought that Emma hoped that the boss would believe that the chemist wanted Emma… (Stiller and Dunbar 101-102)

Correctly answering the questions requires that one be able to entertain a proposition with the relevant order of mental state attributions. Thus, to answer the final question correctly, one must believe that "Jenny thought that Emma hoped that the boss would believe that the chemist wanted Emma." The relevant proposition expresses a fourth order mental state. If one believes the proposition, one has a fifth order mental state. Summarizing a range of experiments that follow the general design sketched above, Dunbar writes: "We have assayed normal adults in an number of separate studies, and it seems that the limits of function for adults is consistently fifth order ("I *believe* that you *suppose* that I *imagine* that you *want* me to *believe* that…") (Dunbar 23). Some individuals appear to be capable of having up to sixth or seventh order mental states (ibid.); however, even these fall below the eighth order mental states I have argued that the expressivist is committed to.

When I first introduced *being for*—the mental state Schroeder posits as part of the expressivist's solution to the scope problem—I suggested that we should understand "being for" as a stand-in for some familiar folk psychological state. Depending on one's preferred version of expressivism, one might take "being for" to stand in for approval, approbation, preference for, etc. If we take Dunbar's research seriously, any expressivist account that attempts to replace *being for* with a familiar folk psychological mental state is doomed to failure. Any such account is committed to humans possessing the capacity to have eighth order mental states composed of multiply embedded folk psychological states. Dunbar's research suggests that we have no such capacity.

There is still a way for the expressivist to sidestep the force of the objection. She can maintain that "being for" picks out a unique mental state that is not identical with any familiar folk psychological state. The expressivist can then hold that *being for* is importantly different from familiar folk psychological states. We can combine *being for* with other folk psychological states to arrive at higher order mental states than we are capable of having with combinations composed of nothing but familiar folk psychological states.

Two remarks about this putative response on behalf of the expressivist. First, if this expressivist response is going to work, we must reify "being for." The trouble is, it is difficult to see any reason to reify "being for" other than to save expressivism. It certainly is not a mental state that I am acquainted with. One might wonder at the plausibility of the expressivist program if it requires, without any pressing explanatory need, that we posit the existence of some mental state of which we are entirely unaware.

Second, supposing that we are willing to reify "being for," it is not methodologically open to the expressivist to insist that there is something special about *being for* that allows us to have eighth order mental states. Remember my earlier methodological critique: we ought not do armchair psychology. This looks like the worst kind of armchair psychology. We are not merely positing the existence of eighth order mental states. Worse yet, we are making claims about unique properties of a certain class of mental states with nothing as evidence except that, were these mental states not to have the properties we want, expressivism would be in trouble. Any naturalistically inclined philosopher should feel deep unease at this proposed move on behalf of the expressivist.

There is, nonetheless, an important kernel of truth in the response I have considered on behalf of the expressivist. We ought not overstate the relevance of the empirical evidence I have cited. At best, it presents *prima facie* reason to think that we can only have sixth or seventh order mental states. It is still *possible* that there is something about moral judgments that allows us to have higher-than-usual order mental states. The case against the expressivist is not yet closed. Nonetheless, the prospects look grim. At the very least, the naturalistically inclined expressivist ought to suspend judgment. While it is possible that cognitive science will vindicate the expressivist's commitments, it certainly has not yet done so. Until that day comes, one can only accept expressivism if one is willing to believe despite a dearth of positive evidence.

Against sentimentalism and constructivism

While expressivism likely constitutes the realist's most popular semantic competitor, various versions of both sentimentalism and constructivism also have widespread support. The realist would do well to offer some reason to be doubtful of both kinds of semantic analyses. In this section I hope to do just that. I will start by considering some work that likely looks largely unrelated. It will take some effort; however, I hope to show that some of the work done regarding normative ethical debates can be made directly relevant to meta-ethical questions.

In a well-known article, "The Secret Joke of Kant's Soul," Joshua Greene draws on his dual process model of moral cognition to launch an attack on deontological moral theories. For the time being, I am relatively uninterested in Greene's arguments. I am, however, very interested in some of the work it has spawned, in particular Kumar and Campbell's reaction. Kumar and Campbell are interested in answering the question, "can empirical evidence be relevant to normative ethical theorizing?" They think that the answer is "yes." They start by articulating a certain style of argument that they label *debunking consistency arguments*. Debunking consistency arguments rely on a familiar normative ethical argumentation strategy: arguments from consistency. Arguments from consistency are familiar and ubiquitous. In brief, one starts with a considered moral judgment, M_1, about some scenario S_1, and compares this moral judgment to some other moral judgment, M_2, about some other scenario, S_2. One then shows that there is no relevant moral difference between S_1 and S_2. It follows that one ought to make the same judgment about S_1 as one makes about S_2. If M_1 is a considered moral judgment and M_1

and M_2 are divergent judgments, one ought defeasibly to reject M_2. Kumar and Campbell offer the following schematic representation of consistency arguments:

> [1] The judgment about case A is a response to F…
> [2] The opposing judgment about similar case B is a response to G…
> [3] The difference between F and G is morally irrelevant.
> Therefore…
> [4] Either the judgment about case A or the judgment about case B is unwarranted. (Kumar and Campbell 322)

Kumar and Campbell think that the argument schema constituted by [1]-[4] offers a promising method by which to make psychological evidence relevant to philosophical ethics. Debunking consistency arguments constitute a species of the genus of consistency arguments. One can offer a debunking consistency argument by introducing a certain kind of empirical evidence to a consistency argument. Kumar and Campbell write:

> Empirical research can be of service here. The research cannot of course tell us whether a proposed difference is morally relevant. What it can tell us is which of the usually many differences between the cases is driving the divergent responses. By constructing cases that differ in only one respect—so-called "minimal pairs"—psychologists can identify the *psychologically efficacious difference*. If the psychologically efficacious difference is not morally relevant, uncontroversially, then the reason we treat the cases differently is not a good reason. (Kumar and Campbell 322)

This passage requires a little unwrapping. The underlying thought is that, by offering subjects minimal pairs, psychologists can determine which features of a scenario are *causally responsible* for the divergence in our moral judgments. Put another way, psychologists can identify the features of a scenario that cause us to make certain types of moral judgments. But we may think that some of these features are morally irrelevant. Thus, we can imagine Roger finding out that the cuteness of kittens and the lack-there-of of pigs is causally responsible for Roger's judgment that it is wrong to kill kittens but permissible to kill pigs. Surely being cute is not a morally relevant property. Nonetheless, it is not entirely clear why making this discovery ought to bother Roger. Facts about what causes Roger to make certain judgments appear to be independent of the truth of

Roger's beliefs. That the cuteness of kittens causes Roger to judge that it is morally impermissible to kill kittens certainly doesn't show that Roger's belief is wrong!

Though the causal account of Roger's psychological states does not show that his beliefs are false, it ought, nonetheless, cast doubt on Roger's belief. The trouble is that, in order for a belief to be doxastically justified, it must have the right basing relationship. That is, for a belief to be doxastically justified it must be non-accidentally related in the appropriate way to the belief's truth-makers. Consider someone who has confirmation that she won the lottery by comparing the numbers on her ticket to the numbers selected in the most recent drawing. Further suppose that she believes that she won the lottery, not because she checked the numbers on her ticket, but because her horoscope said that she would win. Were she not to have seen that her numbers matched the numbers picked in the most recent drawing, she *still* would have believed that she won the lottery. Though she has *justification for her belief*, in the form of seeing that the numbers on her ticket match the winning numbers, her *belief is not justified*—she believes that she has won the lottery for the wrong reasons.

Let us return to Roger. We know that Roger's belief that *it is impermissible to kill kittens* is caused by the fact that the kittens are cute. Furthermore, we do not tend to think that being cute is morally relevant. When we learn that Roger's belief about the impermissibility of killing kittens is caused by the cuteness of kittens, we learn something about the basis for Roger's belief. Roger's belief that *killing kittens is impermissible* may be true; however, it is not doxastically justified. Unless Roger can point to a morally legitimate grounding for his belief, he ought to reject it.

A bit more needs to be said to legitimate this style of argument. When Kumar and Campbell present their original argument schema, the comparison between distinct moral judgments plays an important role. In the sketch I have offered, the comparison to Roger's belief that *it is morally permissible to kill pigs* has, thus far, played no role. In part, the reason is that empirical evidence regarding the proximate cause of distinct moral judgments closes the question of which moral judgment needs to be rejected. In the original consistency argument all that was shown was that *one* of the moral judgments needed to be revised. Empirical evidence regarding the basing of moral beliefs allows us to identify which of the moral judgments has gone wrong. In debunking consistency arguments, the comparison plays an important role in producing the relevant empirical evidence.

The argument rested on the following crucial claim: Roger's belief that *it is impermissible to kill kittens* is caused by the cuteness of the kittens. This claim is, however, clearly absurd. Whatever the correct causal story ends up being, it will certainly involve more than a single property of kittens. The comparative claim allows us to narrow down the causal story. It's not that the cuteness of kittens caused Roger's belief that *it's morally impermissible to kill kittens*. Instead, the cuteness of kittens plays a role in the best causal explanation of the *difference* between Roger's moral judgments.

In a moment I will attempt to link up debunking consistency arguments with meta-ethical considerations; however, before I do so, I need to say a bit about the method of moral semantics. When doing semantics, the goal of the philosopher is to provide jointly necessary and sufficient conditions for the application of a term. If we have given a successful semantic account of, e.g., "permissible," then we will be in possession of a set of conditions such that all and only actions that are permissible fulfill the set of conditions. How does a philosopher determine this set of jointly necessary and sufficient conditions? The thought experiment is the philosopher's primary tool. In offering a thought experiment we attempt to provide a counter-example to some putative analysis. To draw on a putative historical example, suppose that someone suggested that the appropriate analysis of "human" was "featherless biped." A plucked chicken fits these conditions but, of course, is not a human. Thus, a plucked chicken constitutes a counter-example to the putative analysis of "human." In face of this counter-example, we probably ought to reject the analysis. Unfortunately, things are not so simple. In face of a putative counter-example, instead of rejecting her preferred analysis, it is always open to a philosopher to accept the counter-intuitive consequence. If one was *really convinced* that "featherless biped" provided the correct analysis of "human," one might go ahead and accept that a plucked chicken is, in fact, a human. This is, of course, a rather hyperbolic example; however, the general point holds. In giving a semantic analysis, the philosopher aims to capture a great majority of our considered judgments. These considered judgments include both general principles, e.g. all people deserve equal moral consideration, and particular judgments, e.g. it would be wrong to kill Sam in order to steal his dissertation. Thus, in giving an analysis of right/wrong, a philosopher is flexibly beholden to common usage. It is unreasonable to demand that a philosopher aim to capture all aspects of standard linguistic practice; nonetheless, if a putative analysis of "wrong" entails that, *necessarily, it is wrong to utter the word "purple,"* we can be pretty sure that the "analysis" has gone off of the rails. Just as we are justified in rejecting an analysis of wrongness that holds that, necessarily, it is wrong to utter "purple," we are justified in rejecting an analysis of wrongness that forces us to reject a significant portion of our considered normative theorizing.

We can now make both empirical evidence and debunking consistency arguments relevant to meta-ethics. In particular, the two are relevant to the semantic programs of the sentimentalist and the constructivist. In the literature on moral psychology, there are, broadly speaking, three distinct types of views regarding the psychological mechanisms responsible for our moral judgments. Some authors have defended an account of moral psychology where our moral judgments are caused by some range of affective states (Prinz and Nichols). Others have defended a dual process model whereby moral judgments are caused by both conative and cognitive processes (Cushman, Young, and Greene). Finally, some cognitive scientists have defended a view whereby affective states are largely irrelevant to the formation of moral judgments (Dwyer, Huebner, Hauser). I take it that it is possible that, tomorrow, we could find out that any one of these accounts is correct. This fact allows me to illustrate a *prima facie* worry for both sentimentalism and constructivism.

The worry is most easily illustrated if we consider some naïve version of subjectivism. Suppose that one prefers an account whereby "x is wrong" just means, "I disapprove of x." Further suppose that the correct account of moral psychology turns out to be one whereby moral judgments *are not* caused by affective states. This looks like trouble for any such naïve subjectivist. Moral judgments are supposed to be reports about affective states; however, in our imagined scenario, moral judgments are not caused by affective states. It would appear that the subjectivist gets the basing relation wrong; moral beliefs are beliefs about affective states but are only accidentally related to affective states. When doing semantics, we treat our considered judgments as relatively fixed data points. One is allowed to give up on some considered judgments; however, this flexibility is limited. If the facts about our moral psychology are as imagined here, we ought to reject naïve subjectivist. The naïve subjectivist is forced to think that all of her moral judgments are not justified. Either naïve subjectivism is false or we are wrong about the appropriate method of doing semantics.

Now that we see the general structure of the criticism, we can generalize it to all versions of sentimentalism and constructivism. The argument has the following form:

1. In order for our moral judgments to be doxastically justified, semantic view S requires moral psychology P.
2. It is an open question if P holds.
3. Therefore, it is an open question if S is the correct account of moral semantics.

Is it the case that (1) is true? The (cultural) relativist thinks that moral judgments are claims about the shared affective states of some group of people. If (cultural) relativism is the correct account of moral semantics then, in order for our moral judgments to be doxastically justified, they had better be non-accidentally related to whatever cognitive systems are responsible for making judgments about the affective states of others. Ideal observer views and constructivism take moral judgments to be claims about counter-factual scenarios. The proponent of an ideal observer view thinks that “x is wrong” means something like, “an omniscient and disinterested observer would disapprove of x.” On a neo-Rawlsian version of constructivism, “x is wrong” means something like, “x is forbidden by the principles endorsed by rational interlocutors behind the veil of ignorance.” According to a neo-Kantian version of constructivism, “x is wrong” might mean something like, “a rational agent would choose not to do x under any circumstance.” By the lights of any of these views, if a moral judgment is going to be doxastically justified, it must be non-accidentally related to whatever cognitive processes are responsible for making counter-factual judgments.

What ought we make of (2)? It is clear that there is no scientific consensus regarding the cognitive causes of moral judgments. It is, nonetheless, open to the philosopher to argue as follows: I have good reason to believe that semantic account S is true; however, for S to be true it must be the case that moral psychology P obtains. Therefore, I have good reason to believe that moral psychology P obtains. While one can argue this way, I think we would be better off not doing so. Remember my methodological criticism of expressivism: the expressivist *can* argue that we have eighth order mental states; however, doing so commits her to the worst kind of armchair psychology. The same can be said here. The cognitivist anti-realist can save her semantic account by positing that such-and-such is true about our moral psychology. But again, this looks methodologically illicit. The history of science gives us good reason to think that gathering empirical data is the only way to confirm empirical theories. While the cognitivist anti-realist can save her preferred account of moral semantics, doing so requires that she jettison this hard won lesson.

Where does this criticism leave the cognitivist anti-realist? It does not give us any reason to think that any particular anti-realist semantic account is *false*. Instead, the argument demands that the anti-realist bide her time. Before we can have good reason to believe any particular cognitivist anti-realist semantic account, we must have settled the debate over the nature of our moral psychology. Until that debate has been settled we should, at best, be agnostic about sentimentalism and constructivism. If my argument is effective, it constitutes only a provisional win for the moral realist.

One final response on behalf of the cognitivist anti-realist needs to be addressed: what reason do we have to think that the argument does not apply equally well to the realist's preferred semantic account? Nothing in (1)-(3) makes reference to sentimentalism or constructivism. One can just as easily plug moral realism into the argument as either constructivism or sentimentalism. If the argument applies to moral realism as well as it applies to cognitivist anti-realism, it gives us no reason to prefer realism over its competitors. Luckily, I do not think that the argument can be made to apply to the realist's semantic account. Consider, again, the argument's first premise: in order for our moral judgments to be doxastically justified, semantic view S requires moral psychology P. Suppose we replace 'S' with 'moral realism.' What moral psychology does moral realism require? I think that there is no unique answer to this question. The realist's semantic account is only committed to the claim that moral judgments are judgments about ontologically robust entities. What kind of moral psychology would we need to have in order for our judgments about ontologically robust entities to be doxastically justified? Any of the contenders will do. Suppose that it turns out that all of our moral judgments are caused by affective states. So long as these affective states are, themselves, non-accidentally related to ontologically robust moral properties, the realist's moral judgments are doxastically justified. This kind of conative moral psychology, dependent on affective states, contrasts with cognitive accounts of moral psychology, where moral judgments are caused by affect free cognitive processes. Dwyer, Huebner, and Hauser endorse an account of moral judgments whereby they are caused by "a small set of implicit rules and structures [that are] responsible for the ubiquitous and apparently unbounded capacity for making moral judgments" (Dwyer, Huebner, Hauser

486). Again, so long as these implicit rules are non-accidentally related to ontologically robust moral properties, there is no problem for the realist. The realist can tell a similar story regarding other accounts of moral psychology. The realist is not committed to saying anything about the precise nature of moral properties or how we are capable of tracking them. The realist is only committed to the descriptively thin claim that moral properties are ontologically robust. This stands in contrast with the commitments of the cognitivist anti-realist. In giving a particular analysis of moral properties, the cognitivist anti-realist commits herself to "moral properties" having such-and-such a nature. She is thereby further committed to a moral psychology whereby our moral judgments are non-accidentally related to the constituents of her analysis. Thus, each cognitivist anti-realist, but not each moral realist, is committed to a particular account of our moral psychology. The realist escapes the force of the criticism; the sentimentalist and constructivist do not.

Conclusion

I have now introduced the realist's three primary semantic competitors: the non-cognitivist, the sentimentalist, and the constructivist. I have provided a handful of arguments both in favor of the realist's semantic theses and against the theses of her anti-realist competitors. I do not take any of the arguments I have presented, nor the concatenation of all of these arguments, to constitute a conclusive case. Instead, my aim in this chapter was to present *prima facie* reasons for preferring the realist's semantic account. If I am correct that the primary motivation for rejecting the realist semantics is a distaste for the realist's ontological commitments paired with a fear of moral nihilism then establishing a *prima facie* case in favor of the realist semantics ought to be enough. Over the course of the next four chapters I will develop a unified argument in support of the ethical non-naturalist's ontological claims. In so far as this argument is successful, the shadow of moral nihilism can be banished, and the road to the acceptance of the realist's semantic theses made open.

CHAPTER THREE:
THE SCIENTIFIC WORLDVIEW

Introduction

In the opening chapter I briefly laid out the meta-ethical terrain. The reader became familiar with moral realism, the view that I aim to defend. The reader was also introduced to the realist's primary semantic competitors: the non-cognitivist, the sentimentalist, and the constructivist. In the second chapter of the dissertation I offered a handful of arguments intended to create a presumption in favor of the realist's semantic theses. For the remainder of the dissertation, I will presume that my project in the second chapter was successful. The error theorist, or nihilist, is the realist's only remaining meta-ethical competitor. The debate between the realist and the error theorist centers on the ontological claim. The realist and error theorist agree on the conditions a property would have to meet in order to be a *moral* property. The realist thinks that some properties meet these conditions; the error theorist disagrees.

At the outset of the dissertation I made a bold promise: I would argue that acceptance of moral realism—in particular, ethical non-naturalism—is rationally required if one accepts a scientific worldview. In an important sense, the first two chapters of the dissertation were nothing more than preliminary stage setting. The first chapter introduced the reader to the topic at hand and the second chapter pushed back against anti-realist semantics just enough to allow us to address the ontological question I am interested in. The real project of the dissertation starts now. The aim of this chapter is to prepare the reader for the arguments I offer in the remainder of the dissertation.

The goal of my dissertation is to demonstrate that, if one is committed to a scientific worldview, one is thereby committed to ethical non-naturalism. In order to judge the success of my project, we first need to know what I mean by "methodological naturalism" and "scientific worldview." Furthermore, if my project is going to be of much interest to the reader, I need to give the reader some reason to think that there's at least a *prima facie* reason to be doubtful that it is possible to offer a successful methodologically naturalist defense of ethical non-naturalism. My goals for this chapter are two-fold. First, I will clarify what commitments I take to be necessary for the acceptance of a scientific worldview as well as lay out the methods of inquiry that characterize the naturalist methodology. Second, I will present the primary reasons philosophers have for being doubtful that a methodologically naturalist defense of ethical non-naturalism will be forthcoming. By the end of this chapter the reader ought to have a more complete understanding of both the challenges I face and the methods I intend to use in overcoming these challenges.

The scientific worldview and methodological naturalism

The scientific worldview

What, then, constitutes acceptance of a scientific worldview?[8] As is often the case, a rough-and-ready answer is easiest to provide in negative terms. Consider your friend who thinks that a raw food diet cures cancer or consider the neighbor who religiously relies on her horoscope to tell her what her day will bring. Maybe you have an aunt who believes in the healing power of crystals, or an uncle who firmly believes in telepathy. I take it that, in the modern day, acceptance of any of these views is incompatible with a scientific worldview. This fact offers to provide a foothold in our attempt to characterize the scientific worldview. What is it about any of the above listed beliefs that makes holding them incompatible with a scientific worldview?

[8] Determining where to draw the distinction between science and pseudo-science is a particularly vexing philosophical puzzle, known as the "demarcation problem." In light of the difficulty of the demarcation problem, I have little hope of being able to offer a sufficient characterization of the scientific worldview. Rather, I will only offer a rough-and-ready description of the scientific worldview. This kind of broad strokes account ought to be sufficient given the goal of the dissertation.

There is an important similarity between all of these views. What mechanism could be behind the healing power of crystals or the predictive power of horoscopes? By the lights of our best scientific theories, proximity to a lump of stone cannot heal ailments and the positions of planets do not reliably co-vary with events in an individual's life. The mechanisms that are supposed to be responsible for any of the above putative phenomena are, to the best of our knowledge, nomologically impossible.

Accepting the existence of entities/phenomena that are, by the lights of our best theories, nomologically impossible is not, however, a mark of the rejection of the scientific worldview. By way of illustration, consider a relatively recent episode in the history of science. In 2011 a research group at a particle accelerator reported that they had discovered particles that travelled faster than the speed of light (Cartlidge). The theory of relativity, one of our best scientific theories, rules out faster than light travel. Despite the fact that their putative findings contradicted one of our most highly confirmed theories, the research group's trouble-shooting was thorough enough for their data to be published. Given the specific circumstances, believing that *a neutrino could move faster than the speed of light* does not seem to be incompatible with a scientific worldview.

The point can, perhaps, be made more powerfully if we consider the phenomenon of theory change. At the end of the day, the appearance of particles that could move faster than the speed of light was a consequence of an experimental artifact involving a poorly calibrated GPS unit (Cartlidge). But this is not always the case when observed phenomena appear to be at odds with our most highly confirmed scientific theories. Theory change is frequently a consequence of apparent incompatibility between a theory and an observation. Given the importance of new observations overturning old theories, were we to hold that having a belief that was incompatible with current scientific theory was constitutive of abandoning a scientific worldview, we would make one of the most important steps in scientific progress incompatible with a scientific worldview. Though this position is logically consistent, it is certainly counter-intuitive.

The above considerations suggest a new way to proceed. We might do well to figure out what is different about these beliefs—the belief in telepathy, the power of healing crystals, the predictive accuracy of horoscopes— and the kind of beliefs that are incompatible with accepted scientific theory but are, nonetheless, compatible with the scientific worldview.

It may seem that there is an obvious difference between the two sets of beliefs. At one point, there was good evidence for the claim that *neutrons can move faster than the speed of light*. Alternatively, one might think that there is no evidence that telepathy exists, that horoscopes are accurate, etc.

While I think this line of inquiry is promising, further revision is needed. The problem is that there *is* evidence for telepathy, the accuracy of horoscopes, etc. An example will help to illustrate. I once knew a massage therapist who built his own massage table. While massaging a client, the therapist's thoughts drifted and he began to think about what he could do to improve his table. Within thirty seconds, the client asked: "So, how do you like your new massage table?" A week later, the exact same thing happened again, this time with a different client. The massage therapist took these two incidents to be evidence of telepathy. Furthermore, I take it that, on the point of these two incidents counting as evidence for telepathy, the massage therapist was correct. This kind of surprising correspondence between the thought of the massage therapist and the thought of his client is exactly the kind of phenomenon that a theory of telepathy would predict. Thus, these two incidents count as evidence for the existence of telepathy. Similar things can be said about the accuracy of horoscopes, the healing power of crystals, and a raw food diet.

This is not, however, to suggest that there is a preponderance of evidence, or even *good* evidence, in favor of telepathy, etc. There are a variety of criticisms one might push regarding the evidence in favor of telepathy. The sample size is tiny, consisting of only two instances. There is no control group to which we can compare the putatively confirming instances. There are significant worries about cognitive bias: does the massage therapist only remember the instances when there was a correspondence between him thinking about his massage table and his client asking about it? It may well be that there are hundreds of disconfirming instances; however, the massage therapist does not remember any of these. Similar points apply to the kinds of evidence one can marshal in favor of the predictive power of horoscopes and the healing powers of crystals and raw food diets.

We are now in a position to pin down the primary difference between those beliefs that are, and those beliefs that are not, compatible with a scientific worldview. The difference is not that some beliefs are compatible with contemporary scientific theory and others are not. Nor is the difference that there is evidence for some beliefs while there is not evidence for others. Consider the above-mentioned criticisms of the evidence one might have in favor of telepathy. Each criticism points out that the evidence we have in favor of a certain class of beliefs fails to live up to various standards that are commonplace in scientific inquiry. Consider all of the beliefs one might have that are incompatible with contemporary scientific theory. What makes some of these beliefs compatible with a scientific worldview while others are not? The answer seems to be that the evidence we can marshal in favor of some beliefs lives up to a certain set of methodological standards. Thus, the evidence scientists had in favor of the claim that *neutrinos can move faster than the speed of light* was the consequence of careful adherence to the methodological standards that govern scientific inquiry. The evidence we have in favor of telepathy, the accuracy of horoscopes, and the healing powers of crystals and raw food diets is not a consequence of careful adherence to the methodological standards that govern scientific inquiry. Much the opposite.

We are finally in a position to state one of the central commitments of the scientific worldview. Acceptance of the scientific worldview requires that one think that the methods of the sciences constitute an epistemically privileged approach to *a posteriori* inquiry.[9] We cannot believe that telepathy is real without abandoning the scientific worldview because the evidence we have for the existence of telepathy was not collected via methods that live up to scientific standards. On the flip side, one could have believed that *neutrinos can move faster than the speed of light* while accepting the scientific worldview because the evidence we had in favor of the claim was gathered using scientifically acceptable methods.

Two further commitments follow from acceptance of the scientific worldview, i.e. from acceptance of the claim that *the methods of the sciences constitute an epistemically privileged approach to* a posteriori *inquiry*. First, acceptance of a scientific worldview demands the rejection of dogma about the *a posteriori*. If the methods of science constitute an epistemically privileged approach to *a posteriori* inquiry, one can only be epistemically responsible if one is willing to believe in line with the verdict of scientific research. Acceptance of a scientific worldview demands open-mindedness.

Second, acceptance of the scientific worldview requires that one think our best scientific theories are (at least) approximately true. If the methods of the sciences constitute an epistemically privileged approach to *a posteriori* inquiry, then we have very good reason to believe that claims supported by scientific research are true. Our best scientific theories are constituted by the conjunction of scientifically well-confirmed

[9] It is not my intention to downplay the very serious epistemic problems that plague any attempt at *a posteriori* knowledge. The underdetermination of theory by evidence, whether instantiated in the problem of induction or in Cartesian skeptical worries, threatens to undermine the evidential status of scientific inquiry. Luckily, for my project to be successful, I do not need to have a solution to the problem of the underdetermination of theory by evidence. My aim is to offer a defense of the following conditional: if one thinks that the methods of science constitute an epistemically privileged approach to *a posteriori* inquiry, then one ought to be an ethical non-naturalist. As is standard when attempting to prove a conditional, I will assume the truth of the antecedent and see what follows. Happily, the contours of my project allow me to assume that we have a solution to the hard epistemic problems associated with *a posteriori* evidence.

claims. Thus, acceptance of the scientific worldview requires us to think that our best scientific theories are (at least) approximately true.

We are now in a position to nicely summarize a necessary condition for the acceptance of the scientific worldview. Just as there is a debate in meta-ethics about the nature and veracity of moral theories, there is a debate in the philosophy of science about the nature and veracity of scientific theories. Moral realism has an analogue in the philosophy of science literature: scientific realism. The scientific realist is committed to the following four claims:

> 1. Theoretical terms in scientific theories (i.e., nonobservational terms) should be thought of as putatively referring expressions; that is, scientific theories should be interpreted "realistically."
>
> 2. Scientific theories, interpreted realistically, are confirmable and in fact are often confirmed as approximately true by ordinary scientific evidence interpreted in accordance with ordinary methodological standards.
>
> 3. The historical progress of mature sciences is largely a matter of successively more accurate approximates to the truth about both observable and unobservable phenomena. Later theories typically build upon the (observational and theoretical) knowledge embodied in previous theories.
>
> 4. The reality which scientific theories describe is largely independent of our thoughts or theoretical commitments. (Boyd 41-42)

I do not intend to argue for the truth of scientific realism. My aim is not to demonstrate that ethical non-naturalism is correct. Much more modestly (though still quite grandiose), I hope to establish the following conditional: commitment to the scientific worldview further commits one to ethical non-naturalism. We are now in a position to further clarify the conditional I intend to argue for. It is my hope to show that *commitment to scientific realism entails commitment to ethical non-naturalism*.

My general strategy will be straightforward. The scientific realist is committed to the (approximate) truth of well-confirmed scientific theories. If I can show that there are good reasons to think that ethical non-naturalism constitutes a well-confirmed scientific theory, I will have thereby demonstrated that acceptance of scientific realism entails commitment to ethical non-naturalism.

Methodological naturalism

Rather ambitiously, I put the title "A Methodologically Naturalist Defense of Ethical Non-naturalism" on my dissertation. In the philosophical literature, "naturalism" has been used to pick out a variety of philosophical positions. Perhaps most commonly, "naturalism" is used to pick out a metaphysical position whereby the only things that exist are the entities posited by our best microphysical theories (Papineau). When "naturalism" is used to pick out a substantive ontological position, naturalism is, unsurprisingly, incompatible with the ethical non-naturalism I take myself to be defending. Importantly, defenders of naturalism, understood as a substantive ontological thesis, take naturalism to be the only philosophical position that takes science seriously. I disagree. Part and parcel of accepting the scientific worldview is the rejection of dogma regarding the ontological nature of our universe. The methods of science provide epistemically privileged access to the *a posteriori*. If one accepts the scientific worldview, one's beliefs about ontological structure ought to follow from the best science we have available. Thus, it is an open question if naturalism, understood as a substantive ontological thesis, is compatible with the scientific worldview.

Call naturalism, understood as a metaphysical thesis, "ontological naturalism." We can contrast ontological naturalism with what I will call "methodological naturalism." I have argued that acceptance of the scientific worldview requires that one think that the methods of the sciences provide epistemically privileged access to the *a posteriori*. Methodological naturalism is the natural consequence of accepting the scientific worldview. The methodological naturalist holds that, regarding questions of ontology, we ought to defer to the methods of the sciences. Thus, if our best scientific theories entail *such-and-such* about the ontological structure of our universe, we ought to believe *such-and-*

such. If our best scientific theories entail that there are emergent properties then, contra ontological naturalism, we ought to believe that there are emergent properties. If we take acceptance of the scientific worldview seriously, we must accept methodological naturalism. It remains an open question if methodological naturalism suggests ontological naturalism. If it does not, we ought to either reject or, minimally, remain agnostic about, ontological naturalism.

In the remaining chapters of the dissertation I will argue that, if we accept methodological naturalism, we ought to accept ethical non-naturalism. Put another way, I will argue that methodological naturalism and ontological naturalism are incompatible; however, providing a successful argument to this end will require a somewhat more in-depth account of methodological naturalism. It is not enough to merely note that, if we accept the scientific worldview, we ought to take the methods of science seriously. If I am going to defend any substantive ontological thesis on methodologically naturalist grounds, we need some characterization of the methods of science. This characterization can then provide a set of criteria to determine whether or not the defense I have offered is genuinely methodologically naturalist.

Offering an exhaustive account of the going theories of confirmation is obviously well beyond the scope of this dissertation. The same can be said about offering an exhaustive account of the methods of science. Instead, I will focus on the issues directly relevant to the arguments I intend to employ. The purpose of this chapter is to prepare the ground for the defense of ethical non-naturalism I will develop over the remaining three chapters. Consequently, I do not need to provide a survey of either confirmation theory or the methods of science. More modestly, I only need to show that the methods I intend to employ in the remainder of the dissertation are acceptable by the standards of methodological naturalism.

The inadequacy of the naïve model of scientific inquiry

Most likely, when one thinks of the methods of the sciences, one thinks of empirical confirmation of theory. On a familiar model of the methods of science, a theory is proposed, the scientist derives some prediction about observable states-of-affairs that would follow were the theory true, then the scientist runs an experiment to see if the observational predictions hold. If the prediction is confirmed, we have further evidence for our theory. If the prediction is not confirmed, we have some reason to think that our theory has gone wrong. Call this the "naïve model of scientific inquiry."

It will likely come as little surprise that the naïve model of scientific inquiry fails to accurately describe the project of the remaining three chapters. I am a philosopher, not a scientist. I am trained to find conceptual connections and to eliminate conceptual ambiguities. Experimental design, procedure, and techniques are all beyond my ken. It seems best to leave the construction and implementation of experiments to those with the relevant skill sets. Consequently, if one thinks that theory confirmation via the methods of the sciences is exhaustively composed of testing theoretical predictions via experimentation, one will think that the arguments in the following chapters fail to live up to the standards of methodological naturalism. Fortunately, it is widely accepted that scientific methods of theory confirmation are significantly more complicated than the naïve model of scientific inquiry suggests.

Two problems for the naïve account of scientific confirmation are particularly relevant to my project. The first problem has to do with the time-independence of evidence. An example will help to illustrate. Suppose that we have a six-sided die. As is generally the case with six sided dice, each face shows an integer between one and six; no integer shows up more than once. Suppose we have the following two competing hypotheses:

1. [Die_{fair}] On a given roll, there is an equal probability of any face of the die landing up.

2. [Die$_{\text{loaded}}$] On a given roll, there is a 50% probability that a face showing a six will land up.

Further suppose we have data regarding ten rolls of the die. In all ten rolls, a face showing a six landed up. Presume, for whatever reason, that we can count out all other hypotheses other than [Die$_{\text{fair}}$] and [Die$_{\text{loaded}}$]. It looks like we have very good evidence for [Die$_{\text{loaded}}$]. If [Die$_{\text{loaded}}$] holds, it is significantly more likely that a face showing a six would be rolled ten times in a row than if [Die$_{\text{fair}}$] were true. Whether [Die$_{\text{fair}}$] and [Die$_{\text{loaded}}$] were formulated before or after we starting rolling dice is unrelated to the extent to which our evidence favors one hypothesis over the other. We can propose hypotheses, make empirical predictions, and then run the tests *or* we can run the tests and only then consider what hypothesis they might support. The temporal relation between forming a hypothesis and gathering data relevant to the hypothesis is unrelated to the extent that the data counts as evidence for a hypothesis. Call this fact the *temporal independence of evidence.*

The *temporal independence of evidence* has important implications for the naïve model of scientific inquiry. On this account, the methods of science involve deriving observational predictions from theories and then testing to see if these predictions hold. This view of scientific confirmation, call it "naïve predictivism," has fallen out of favor. The problem is that the naïve predictivist appears to be committed to the denial of the temporal independence of evidence. On the naïve predictivist picture one *first* formulates a theory and *then* one looks for confirmation. In light of the temporal independence of evidence, a theory can be well confirmed even if no one has run an experiment specifically designed to test the theory.

It is important to be clear about the strength of the claim I take myself to be making in the preceding paragraph. There may be an important sense in which a theory that makes a novel prediction, which is subsequently confirmed, is more virtuous than a theory that is merely designed to account for evidence we have already gathered. This issue is of some debate and I do not intend to take a side. For my purposes here, it is sufficient to note that a theory can be well confirmed even if no one has run an experiment specifically designed to test the theory. The arguments I offer in the following three chapters can live up to the standards set by methodological naturalism even though I have failed to offer any novel experimental evidence.

There is a second, more worrisome, problem with the naïve model of scientific inquiry. By the lights of the naïve model, the scientist derives an observational prediction from a theory then runs an experiment designed to bring about the predicted observations. If the prediction holds, we have good reason to believe the theory. If the prediction does not hold, we have good reason to reject the theory. The second problem with the naïve account goes by the name *the Quine-Duhem problem* or the *Quine-Duhem thesis*. Again, the problem is most easily illustrated by example.

Reconsider the experiments that putatively demonstrated that neutrinos could move faster than the speed of light. On face, these observations seem to be incompatible with the theory of relativity; however, a little reflection on the process by which the experimental results were rejected will demonstrate that this is not the case. The experiment results that putatively showed that neutrinos could move faster than the speed of light were rejected after an error in a GPS unit was discovered. This story of experimental trouble shooting highlights an important point: the theory of relativity may, all on its own, predict that neutrinos cannot move faster than the speed of light; however, the theory of relativity, by itself, has *nothing* to say about the kind of observations scientists gather from particle accelerators. Instead, one must conjoin the theory of relativity with a vast range of auxiliary hypotheses in order to make empirical predictions. In the present case, the auxiliary hypotheses include some set of claims regarding the relation that holds between the measurements of a GPS unit and the speed of particles.

The reader might, reasonably, wonder what the term "auxiliary hypothesis" is supposed to pick out. I am afraid that I do not have a rigorous answer; however, an example should help illustrate. Consider, again, the putative faster-than-light travel of neutrinos. These experimental results were not the result of tests that aimed to either confirm or disconfirm the theory of relativity. Nonetheless, for the remainder of this paragraph, let us pretend that the goal in running particle accelerators was to test the theory of relativity. Were this the case, we would have an instance of an experiment designed to test a particular theory—in this case, the theory of relativity. Let R denote the set of propositions constitutive of the theory of relativity. Any hypothesis not in R that plays a role in establishing probabilistic relationships between R and observation statements is an auxiliary hypothesis. While also admitting his inability to offer a rigorous definition, Thagard offers the following related definition:

> The explanation of facts *F* by a theory *T* requires a set of given conditions *C* and also a set of auxiliary hypotheses *A*…. An *auxiliary hypothesis* is a statement, not part of the original theory, which is assumed in order to help explain one element of *F* or a small fraction of the elements of *F*. (Thagard 86)

I take the definition I have offered and the definition Thagard offers to be co-extensive. If some auxiliary hypothesis, in conjunction with a theory, is going to explain some fact, the auxiliary hypothesis must play a role in establishing some probabilistic relationship between the theory and the fact it is attempting to explain.

Having offered some account of what is meant by "auxiliary hypothesis," let us return to the *Quine-Duhem* thesis. Once we see that the theory of relativity, by itself, is not sufficient to make predictions about observations, we further see that no experimental evidence, by itself, counts against the theory of relativity. Consider a counter-factual case where scientists gathered observational data suggesting that neutrinos can move faster than the speed of light and no flaws in the experimental design were discovered. Must we then reject the theory of relativity? The answer is "no." For example, we might decide that something has gone wrong with the auxiliary hypotheses underlying our use of global positioning satellites, e.g. maybe atmospheric interference makes it impossible to use satellites to make position measurements with the required degree of accuracy.

The Quine-Duhem thesis holds that it is impossible to test any given theory in isolation, for no single theory makes observational predictions. One can only derive observational consequences from theory in conjunction with auxiliary hypotheses. This presents two distinct problems for the naïve model of scientific inquiry. First, the relationship between experimentation and theory rejection is not as straightforward as it might seem. Given that no single theory makes observational predictions, if our experiment does not turn out as expected, we have a wide array of options open to us. Had there been no GPS error in the experiments that putatively showed neutrinos moving faster than the speed of light, we would not have been forced to reject the theory of relativity. We could have, instead, chosen to reject any of the auxiliary hypotheses that, when conjoined with the theory of relativity, lead to incompatibility with experimental results.

Second, theory confirmation is not as straightforward as the naïve model of scientific inquiry suggests. Suppose some concatenation of theories entails an observational prediction. Furthermore, this observational prediction is experimentally confirmed. The observational prediction in question was the consequence, not of a single theory, but of a conjunction of a theory and some set of auxiliary hypotheses. Consequently, the experimental findings do not *merely* confirm a single theory. Instead, theory confirmation is holistic. When an experimental finding confirms a theoretical prediction, every auxiliary hypothesis included in the conjunction required to make the observational prediction is confirmed.

The temporal independence of evidence and the Quine-Duhem thesis show that the naïve model of scientific inquiry is inadequate. In light of the Quine-Duhem thesis and the temporal independence of evidence, I am now in a position to add to the naïve model of scientific inquiry. In what follows, I will sketch one important method by which theories are confirmed. The arguments for ethical non-naturalism I develop in the remaining chapters of the dissertation will live up to the standards set by methodological naturalism in so far as they correspond to the method of theory confirmation I outline in the remainder of this section.

Inference to the best explanation

It is widely accepted that one of the primary aims of the sciences is to provide explanations of observed phenomena. There is, unfortunately, significant disagreement about what constitutes an explanation and even more disagreement about what constitutes a successful explanation. This is not, however, to suggest that there is no consensus. This much, at least, everyone appears to agree on. First, scientific theories are intimately connected to explanations. Theories, or more accurately conjunctions of theories and auxiliary hypotheses, are in the business of offering explanations. If one wants an explanation of, e.g. the population dynamics of moose and wolves on Isle Royale, one looks to theories of ecology. Alternatively, if one wants an explanation of heat transfer, one looks to thermodynamics. Scientific theories are not scientific explanations; however, explanations can be derived from scientific theories.

Second, a scientific theory can be confirmed in so far as it offers to provide a best explanation (or is the member of a set of scientific theories the conjunction of which provides a best explanation). Explanations can be better or worse depending on the extent to which the theories from which they are derived possess the super-empirical virtues. Again, there is significant disagreement regarding what counts as a super-empirical virtue. Nonetheless, as before, there are pockets of consensus. It is widely accepted that the *simplicity* and *consilience* of a theory count in favor of an explanation.

Consilience is a measure of the range of phenomenon a theory can explain. *Ceteris paribus*, an explanation is better if it can be derived from a more consilient theory (or set of theories). Thagard offers the following account of comparative consilience:

> [L]et FT_i be the set of classes of facts explained by [a set of theories] T_i… [W]e can choose between different definitions of comparative consilience: (1) T_1 is more consilient than T_2 if and only if the cardinality of FT_1 is greater than cardinality of FT_2; or (2) T_1 is more consilient than T_2 if and only if FT_2 is a proper subset of FT_1. These definitions are not equivalent, because FT_1 might be much larger than FT_2, while at the same time there are a few elements of FT_2 that are not in FT_1. In other words, it is possible that T_1 explains many more classes of facts than T_2, but that there are still some facts that only T_2 explains. In cases where these two definitions do not coincide, decisions concerning the best explanation must be made according to what theory explains the most important facts, or on the basis of other criteria… (Thagard 79-80)

Put in a more rough and ready way: if two explanations are in competition, we ought to prefer whichever explanation is derived from a theory, the conjunction of which alongside some set of auxiliary hypotheses, offers to explain the most phenomena.

There is also wide agreement that simpler explanations are, *ceteris paribus*, better explanations. Unfortunately, "simplicity" is not, itself, easy to make sense of. There are a variety of distinct ways one can think about simplicity. One might think about simplicity in terms of syntax and one can think about syntactic simplicity in a variety of ways. For example, one might measure syntactic simplicity in terms of the *length* of an explanation. Thus, an explanation that requires fewer lines/sentences/syntactic clauses is syntactically simpler than a longer explanation. Alternatively, one might think about syntactic simplicity in terms of the number of variables an explanation requires. An explanation the syntax of which only involves five variables is, on this view, syntactically simpler than an explanation the syntax of which involves six variables. On a third account of syntactic simplicity, the simplicity of an explanation can be measured by the number of distinct auxiliary hypotheses required for the explanation.

There is good reason to think that considerations of syntactic simplicity—when appropriately cashed out—are important. A theory can always be made to fit the evidence so long as one is willing to add *ad hoc* auxiliary hypotheses to cover special cases. Ptolemic astronomy offers the paradigm instance of this kind of special-case-pleading. Ptolemaic astronomy held that the heavenly bodies were in circular orbits around the earth. Unsurprisingly, this model of the solar system failed by itself to offer accurate predictions of the positions of the planets in the night sky. Particularly troublesome for the account was retrograde motion, a phenomenon where planets appear temporarily to move in the opposite direction of their normal orbit. Ptolemy's "solution" came in the form of epicycles—essentially a small loop added to a planet's otherwise circular orbit. With the addition of enough epicycles, Ptolemaic astronomy was a strikingly predictively successful theory. The addition of each epicycle to Ptolemaic astronomy improved predictive accuracy at the cost of increased syntactic clutter. While the point is nicely illustrated by Ptolemaic astronomy, it is entirely general. One can always improve the predictive accuracy (as well as the consilience) of a theory by adding *ad hoc* auxiliary hypotheses; however, doing so will always require the addition of syntactic clutter. Thus, some account of syntactic simplicity as a super-empirical virtue may be required to help rule out the illicit addition of *ad hoc* auxiliary hypotheses.

Nonetheless, my arguments in the remainder of the dissertation will, by and large, ignore considerations of syntactic simplicity. Instead, I will focus on a second kind of simplicity: ontological simplicity. An explanation that requires we posit fewer ontological entities is, *ceteris paribus*, the better explanation. By way of illustration, consider the debate in the philosophy of mind between the substance dualist and the materialist. The materialist about the mind thinks that minds are composed of nothing over-and-above the physical. The substance dualist disagrees. The substance dualist holds that there is some kind of ontological entity, distinct from physical matter, of which minds are composed. The materialist is committed only to the existence of physical matter. The substance dualist is committed to the existence of both physical matter and mental substance. Were the two theories otherwise equally virtuous, considerations of ontological simplicity suggest that we should prefer materialism.

While there is much more that can be said about the super-empirical virtues, for the purposes of this dissertation, we have reached the point of diminishing returns. My arguments for ethical non-naturalism will focus on the super-empirical virtues of consilience and simplicity. I will, eventually, argue that ethical non-naturalism is more consilient than its competitors. I will also argue that considerations of ontological simplicity do not count against the ethical non-naturalist. This is, however, largely foreshadowing. For now, my aims are much more limited.

The aim of this section was to sketch some of the methods of inquiry condoned by methodological naturalism. We can then use this sketch to determine whether the arguments I offer in the remainder of the dissertation constitute a methodologically naturalist approach to meta-ethics. The original cause for concern was that, in the remaining chapters, I will not offer any novel experimental results. On at least one understanding of the scientific method, this fact is enough, from the perspective of methodological naturalism, to keep my arguments from being respectable. Luckily for me, both the *temporal independence of evidence* and the *Quine-Duhem thesis* undermine the view that scientifically respectable evidence can only come from novel experiments. Having briefly discussed the super-empirical virtues, I am now in a position to offer a sketch of a method of theory confirmation that is both ubiquitous throughout the varied fields of science and does not require novel experimental data.

Inference to the best explanation (henceforth, IBE)—also known as "abduction"—constitutes one of the primary methods by which scientific theories are confirmed. One of the roles of a scientific theory is, along with some set of auxiliary hypotheses, to explain some range of phenomena. Different theories, in addition to their corresponding auxiliary hypotheses, offer competing explanations of the same phenomena, e.g. both neo-Darwinian evolutionary theory and the theory of intelligent design attempt to explain a range of facts regarding the phenotypic and genotypic traits of organisms. The theories are incompatible; they can't both be right. Presuming that both theories can offer an explanation of some of the phenomena in question, we rely on the super-empirical virtues to determine which explanation is better. If neo-

Darwinian evolutionary theory offers a better explanation, in the sense that neo-Darwinian evolutionary theory possesses a preponderance of the super-empirical virtues, then we ought to believe neo-Darwinian theory instead of its competitor, intelligent design theory. The same can be said of any given set of explanations, $\{E_1, E_2, \ldots, E_n\}$, competing to explain some set of phenomena, $\{p_1, p_2, \ldots, p_n\}$. Each explanation will be derived from some combination of a theory and auxiliary hypotheses, $\{TA_1, TA_2, \ldots, TA_n\}$. If E_1 is the best explanation of the bunch, and E_1 is derived from TA_1, then we have reason to believe TA_1 is true. If TA_1 is true, and TA_1 posits the existence of some kind of entity, then we have good reason to believe that said entity exists. IBE gives us a way to do ontology that is compatible with methodological naturalism.

Importantly, the *temporal independence of evidence* and the *Quine-Duhem thesis* both play an important role in IBE. Explanations are always explanations of observed phenomena; where one lacks observational data, one has nothing to explain. Consequently, IBE always relies on observational data that has *already* been gathered. Whereas the naïve model of scientific inquiry appeared to demand that we first formulate hypotheses and then test them, IBE reverses the temporal relationship between hypotheses and observational evidence. IBE is one of the primary methods of confirming a scientific theory and, by its very nature, does not require the presentation of novel experimental data.

The Quine-Duhem thesis highlighted the fact that confirmation is holistic. No single theory, by itself, makes observational predictions. Instead, one needs a theory alongside a range of auxiliary hypotheses. The confirmation of one's empirical predictions does not just count as evidence for one's theory, it also counts as evidence for one's auxiliary hypotheses. The same holds for IBE. A theory cannot, by itself, offer an explanation of some phenomenon. Instead, explanations are composed of the conjunction of a theory and some set of auxiliary hypotheses. Consequently, if the best explanation E_1 is derived from the conjunction of some theory, T_1, and some set of auxiliary hypotheses, $\{A_1, A_2, \ldots, A_n\}$, then one has evidence that both T_1 and the members of $\{A_1, A_2, \ldots, A_n\}$ are correct. Both the temporal independence of evidence and the Quine-Duhem thesis play important structural roles in IBE. The very reasons we had for thinking that the naïve model of scientific inquiry failed to offer a correct account of theory confirmation are accounted for in the structure of abductive reasoning.

One further point needs to be made before proceeding. It is not always the case that IBE proceeds as a careful examination of the super-empirical virtues of competing explanations. "Best" is a comparative notion. Consequently, there is a surefire way for some explanation to count as the best explanation. If there is only one explanation available, *a fortiori* that explanation is the best explanation. This style of argument, whereby one shows that such-and-such an explanation is the only explanation available, is known as a *no miracles* argument. Supposing that there is only one explanation of phenomenon P available, were this explanation false, P would be unexplainable—a miracle. Determining which explanation is most virtuous is no easy task, though philosophers sometimes seem to write as if this were not the case. In light of the difficulty of determining which explanation is most virtuous, IBE is most effective when one can show that there is only a single adequate explanation.

The methodologically naturalist challenge to ethical non-naturalism

The received view in contemporary meta-ethics appears to be that acceptance of a scientific worldview is incompatible with ethical non-naturalism. Even those proponents of ethical non-naturalism who think that the view is compatible with the scientific worldview are skeptical that any methodologically naturalist defense of ethical non-naturalism will be forthcoming (see, e.g., Shafer-Landau "Moral and Theological Realism: The Explanatory Argument"). The remaining four chapters of the dissertation are aimed at demonstrating that acceptance of the scientific worldview gives one good reason to accept ethical non-naturalism. Before offering my methodologically naturalist defense of ethical non-naturalism, I will sketch some of the reasons one might have for thinking that the scientific worldview and ethical non-naturalism are incompatible.

Some rough and ready worries

Before presenting two of the better-known arguments for the conclusion that ethical non-naturalism is incompatible with the scientific worldview, I will briefly sketch some rough and ready considerations that might motivate one to think that the two positions are incompatible. I don't intend for anything I say in this section to constitute an argument against the compatibility of ethical non-naturalism and the scientific worldview. My aim is merely to point to some considerations that might lead one to suspect that some such argument will be forthcoming.

A brief look at the history of science may be enough to cause concern about the compatibility of the scientific worldview and ethical non-naturalism. Beliefs about vital humors, biological *teloi*, and various supernatural powers were, historically, prevalent. The advance of science has progressively demonstrated that these mysterious properties are non-existent. Science eliminates the supernatural and the mysterious and replaces it with the natural and the understandable. Non-natural ethical properties may strike one as a likely candidate for elimination via scientific progress. By hypothesis, non-natural ethical properties are not describable via the laws of physics. Non-natural ethical properties cannot be described in the language of our most successful scientific theories nor can they be straightforwardly causally influenced by the physical. If scientific progress eliminates mysterious properties, then non-natural ethical properties should be next up on the chopping block.

Mackie has, famously, marshaled similar considerations against ethical non-naturalism, though his argument lacks the historical trappings of the above considerations. In Mackie's words,

> [i]f there were objective values, then they would be entities or qualities or relations of a very strange sort, utterly different from anything else in the universe… An objective good would… [have] to-be-pursuedness somehow built into it. Similarly, if there were objective principles of right and wrong, any wrong (possible) course of action would have not-to-be-doneness somehow built into it.(Mackie 76-77)

Mackie thinks that the putative queerness of moral properties constitutes a sufficient reason to reject moral realism. Mackie's rejection of ethical non-naturalism on these grounds is a consequence of his acceptance of ontological naturalism. I have argued that ontological naturalism constitutes a metaphysical dogma. We only ought to accept ontological naturalism if it follows from methodological naturalism. This is an open question. For the time being, the ethical non-naturalist has nothing to fear from Mackie's argument from queerness. Nonetheless, with a minor retrofit, we can turn the argument from queerness into a more pressing challenge.

The moral anti-realist can offer the following meta-inductive consideration: moral properties would be unlike any of the properties science has yet to discover; therefore, we ought to doubt that any such properties will be identified via scientifically acceptable methods. The previous considerations I offered that might make one think that methodological naturalism and ethical non-naturalism are incompatible relied on a tenuous analogy between supernatural properties and moral properties. This retrofit of Mackie's argument from queerness requires no such analogy. The mere recognition that non-natural ethical properties look unlike familiar scientific properties might be enough to raise doubts about the compatibility of ethical non-naturalism and methodological naturalism.

Before moving on, I'd like to suggest one further set of considerations that might make one think that a successful argument for the incompatibility of ethical non-naturalism and methodological naturalism will be forthcoming. There is a well-known dictum in ethical theory: one cannot derive an "ought" from an "is." David Hume originally presented the view that one cannot derive normative conclusions from descriptive premises:

> In every system of morality, which I have hitherto met with, I have always remarked, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surprised to find that, instead of the usual copulations of propositions, *is*, and *is not*, I meet with no proposition that is not connected with an *ought*, or an *ought not*. This change is imperceptible; but it is, however, of the last consequence. For as this *ought*, or *ought not*, expresses some new relation or affirmation, it is necessary that it should be observed and explained; and at the same time a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it. (Hume 469)

It is a matter of some debate whether it is possible to reach a normative conclusion from purely descriptive premises; however, the impossibility of doing so is at least *prima facie* plausible. It appears that the sciences are in the business of offering purely descriptive sentences. One does not find much moral language in physics, chemistry, or biology journals. If there is an is/ought gap, then it is difficult to imagine how one might offer a methodologically naturalist defense of ethical non-naturalism. The data of the sciences do not speak to questions of morality.

I do not intend for any of the above considerations to constitute a powerful argument against the possibility of providing a methodologically naturalist defense of ethical non-naturalism. Instead, I hope that they help to make clear the enormity of the task on which I will embark in the remaining three chapters of the dissertation. However one wants to cash it out, there appears to be an important difference between the subject matter of the sciences and the subject matter of meta-ethics. Given that the methods of each science appear to be uniquely suited to the study of its respective domain and that meta-ethics does not appear to fall under any of the relevant domains, it ought to appear that the odds of providing a methodologically naturalist defense of ethical non-naturalism are long.

Harman's critique

Gilbert Harman is responsible for perhaps the most influential argument that aims to demonstrate that methodological naturalism and ethical non-naturalism are incompatible. Harman's argument relies heavily on a putative disanalogy. He asks us to consider two distinct cases. In the first case, "You see some children pour gasoline on a cat and ignite it… [Y]ou make a moral judgment immediately and without conscious reasoning … that the children are wrong to set the cat on fire…" (334-335). Harman offers a second case, by way of contrast:

> Consider a physicist making an observation to test a scientific theory. Seeing a vapor trail in a cloud chamber, he thinks, "There goes a proton." Let us suppose that this is an observation in the relevant sense, namely, an immediate judgment made in response to the situation without any conscious reasoning having taken place. (334)

As Harman notes, there is no such thing as pure observation. All observation is theory dependent. In virtue of our (presumably tacit) moral theory, we can be said to have observed the wrongness of torching the cat. Analogously, in virtue of his physical theory, the physicist can be said to have observed a proton. Nonetheless, Harman thinks that the physicist's observation counts in favor of the ontologically robust existence of protons whereas our observation of the wrongness of burning a cat does not count in favor of the ontologically robust existence of rightness and wrongness. What, then, is the relevant disanalogy?

Harman notes that, "[t]he observation of an event can provide observational evidence for or against a scientific theory in the sense that the truth of an observation can be relevant to a reasonable explanation of why that observation was made" (335). This ought to sound familiar. Inference to the best explanation is one of the primary methods of scientific theory confirmation. We have reason to believe the theory that best explains some observed phenomena. If the best explanation of the physicist's observation follows from a theory that posits the existence of protons, we have good reason to believe that protons exist. Analogously, if the best explanation of our observation that it is wrong to burn a cat follows from a theory that posits the existence of non-natural moral properties, we have good reason to believe that non-natural moral properties exist. However, Harman argues that, while the best explanation of the observed proton requires that we posit the existence of a proton, the best explanation of our moral observation does not require that we posit the existence of any non-natural moral properties:

> A moral observation does not seem, in the same sense, to be observational evidence for or against any moral theory, since the truth or falsity of the moral observation seems to be completely irrelevant to any reasonable explanation of why that observation was made. The fact that an observation of an event was made at the time it was made is evidence not only about the observer but also about the physical facts. The fact that you made a particular moral observation when you did does not seem to be evidence about moral facts, only evidence about you and your moral sensibility. Facts about protons can affect what you observe, since a proton passing through the cloud chamber can cause a vapor trail that reflects light to your eye in a way that, given your scientific training and psychological set, leads you to judge that what you see is a proton. But there does not seem to be any way in which the actual rightness or wrongness of a given situation can have any effect on your perceptual apparatus. In this respect, ethics seems to differ from science. (335)

Harman further elaborates:

> In the moral case, your making your observation does not seem to be evidence for the relevant moral principle because that principle does not seem to help explain your observation. The explanatory chain from principle to observation seems to be broken in morality. The moral principle may "explain" why it is wrong for the children to set the cat on fire. But the wrongness of that act does not appear to help explain the act, which you observe, itself. The explanatory chain appears to be broken in such a way that neither the moral principle nor the wrongness of the act can help explain why you observe what you observe. (336)

In the case where a physicist observes a proton, it appears that the best explanation requires that we posit the existence of a proton. Were there nothing in the external world causally responsible for our experience as of a proton, our experiences would be inexplicable: a miracle. Our best theories regarding the external world tell us that the object causally responsible for our observation is a proton. Consider the contrast class: our observation of the wrongness of burning a cat. Well-confirmed scientific theory provides all of the resources necessary to explain our visual experience. Some explanation is still needed of our experience of the *wrongness of burning a cat*. But as Harman notes, psychological theory can do all of the needed work. Furthermore, everyone needs to offer some psychological theory that relates our visual experiences to our experiences as of wrongness. Once we have such a theory, it seems that the conjunction of physical and psychological theory can offer a simple and consilient explanation of our observation of the wrongness of burning a cat. Positing the existence of non-natural moral properties doesn't seem to add anything to the explanation; however, it does make the explanation less ontologically simple. Thus, it is implausible to believe that the best explanation of any moral observation will involve reference to non-natural moral properties.

Harman's argument is, *prima facie*, devastating to the project of offering a methodologically naturalist defense of ethical non-naturalism. My aim is to offer an abductive argument for the existence of non-natural ethical properties. IBE only works if we have some observation in need of explanation. But if Harman is correct, the best explanation of moral observations will not involve reference to non-natural moral properties. Thus, we have good reason to be pessimistic about the prospects for a methodologically naturalist defense of ethical non-naturalism.

A Rejoinder to Harman and Leiter's Response

Harman's argument is not as straightforwardly problematic as it may first appear. Notably, not all explanations are explanations of *such-and-such an observation*, where this is read as referring to some set of experiences or mental states. If every explanation is of some experiential state, it is difficult to see how explanations will ever go beyond the psychological. Consider a rather mundane example. If my car does not start, I am not interested in explaining the set of experiences constitutive of my experience of my car not starting. Instead, I am interested in explaining *why my car does not start*. Consequently, in the explanatory game, at least some observations get to count, defeasibly, as veridical. It is open to the ethical non-naturalist to hold that the appropriate explanandum is not the fact that *we had an experience as of the wrongness of burning a cat*, but is, instead, the fact that *it is wrong to burn a cat*.[10] On this latter approach, prospects for the ethical non-naturalist look more hopeful. Presumably some set of moral principles best explains why *it is wrong to burn a cat* (Shafer-Landau "Moral and Theological Realism: The Explanatory Argument"). If the ethical non-naturalist can win the semantic debate, then a moral theory constituted by the moral principles in question posits the existence of non-natural moral properties and, consequently, we have good reason to believe that there are non-natural moral properties.

Brian Leiter has pushed back against this argumentative response on behalf of the ethical non-naturalist. Leiter relies heavily on the super-empirical virtue of consilience to make his point. He writes:

> Applying Thagard's criteria yields a standard attack on the status of various putatively real facts, an attack based on what I will call *the problem of explanatory narrowness* (PEN). A property suffers from PEN if its explanatory role is too peculiar or narrow, that is, if it only explains one class of phenomena to which it seems too neatly tailored. Real explanatory facts, Thagard's criteria suggest, must have some degree of extra consilience. Properties that "explain," but suffer from PEN, are not "real" properties. (Leiter 82)

[10] This is not a point that is lost on Harman.

Leiter then suggests that we have reason to believe that explanations of our moral observations that posit the existence of moral properties suffer from the problem of explanatory narrowness:

> We can see that … [moral explanations] are in fact inferior to … [naturalistic explanations] by attending again to Thagard's … criteria… [C]onsider consilience… [Naturalistic explanations] will always explain more than … [moral explanations] do. This is because the mechanisms employed by … [naturalistic explanations] explain much more than just the class of "moral" phenomena (e.g., moral beliefs and observations), while … [moral explanations] will only be able to explain the moral phenomena. This should hardly be surprising: after all, … [naturalistic explanations] were generally proffered as accounts of other phenomena… (Leiter 88)

Presuming that moral theory can only explain the class of putative moral facts, moral theory does not look particularly consilient. One might think that the supposed lack of consilience of moral theory is not much of a problem. After all, we are out to explain *moral facts*. Surely nothing but moral theory can do the trick. The problem is that we allowed moral facts into the domain of explanadum because of the theory ladeness of observation. The conjunction of our moral theory and the theory ladeness of observation allows the ethical non-naturalist to posit the existence of a domain of moral facts. However, the veracity of our moral observations is defeasible. Consider an analogy to observations of witches. If one has a theory of witches and observation is theory laden, one can putatively observe witches; however, the explanatory failures of witch theory cast the theory in to doubt, which in turn gives us reason to doubt the veracity of our putative observations of witches. The same can be said regarding moral theory and putative moral observations. The comparative lack of virtue of the moral theory that explains moral facts, as compared to the physical, biological, and psychological theory that attempts to explain our experience as of moral properties, gives us reason to doubt both the moral theory and the claim that we have veridical moral observations.

Conclusion

In the first half of this chapter I defined methodological naturalism and offered a very brief overview of the methods I take to be endorsed by the view. If the defense of ethical non-naturalism I am about to offer is going to live up to the standard I have set for it, it must rely on these methods. In the second half of the chapter I briefly motivated and sketched arguments for the view that no methodologically naturalist defense of ethical non-naturalism would be forthcoming. Conspicuously absent from this chapter is any attempt to respond to these arguments. Indeed, I have not even offered a sketch of the most prevalent responses ethical non-naturalists have given to the arguments of Harman and Leiter. There are two reasons for my silence. Most importantly, I see no reason to argue that, in principle, it is possible to offer a methodologically naturalist defense of ethical non-naturalism. If I can successfully give such a defense I will have shown, *a fortiori*, that such a defense can be given. If I cannot give such a defense, knowing that it is in principle possible to do so will come as cold comfort. To the best of my knowledge, the defense of ethical non-naturalism I am about to offer is unique in that it is the only such defense that draws on established scientific theory. I take it that the received view in meta-ethics is that the ethical non-naturalist has yet to provide a satisfactory methodologically naturalist defense against ethical non-naturalism. Insofar as this assessment is accurate, the only reason one might have for surveying attempted responses would be to show where they went wrong. There will be some of this in the fifth chapter; however, given that I intend to take an entirely novel approach, it does not appear that there is much to be gained by rehashing old arguments. Moreover, even if the received view is inaccurate and the ethical non-naturalist has provided a satisfactory methodologically naturalist defense of her view, there still seems to be little reason to rehash old territory. In so far as the arguments I offer in the remaining three chapters are both novel and successful, we will have even better reason to think that acceptance of the scientific worldview further commits one to ethical non-naturalism.

CHAPTER FOUR:
A METHOD FOR IDENTIFYING NON-NATURAL PROPERTIES

Introduction

The goal of my dissertation is to provide a methodologically naturalist defense of ethical non-naturalism. Ethical non-naturalism is characterized by a commitment to the existence of non-natural moral facts. A fact is non-natural just in case it is not token-token identical with any set of microphysical facts. By the end of this chapter I aspire to have identified an argument schema such that, if a moral predicate successfully instantiates the schema, we have good reason to believe that there are non-natural moral facts. In the following chapter I will draw on this argument schema in an attempt to vindicate ethical non-naturalism.

A reminder: my definition of non-naturalism

Given that the aim of this chapter is to provide an argument schema capable of identifying non-natural properties, it is important to remind the reader that I am working with a non-standard definition of non-naturalism. In the first chapter I noted that, in philosophy at large, there appears to be no widely accepted way of drawing the distinction between naturalism and non-naturalism. I then considered how meta-ethicists use the two terms. I found that Brink and Shafer-Landau have apparently identical views regarding the metaphysics of moral properties; however, Brink is a highly regarded naturalist and Shafer-Landau is a highly regarded non-naturalist. No definition of "non-natural" would be forthcoming from an examination of how meta-ethicists use the term.

I then suggested that, however we define "non-natural," we need to be able to make sense of the worry that non-natural properties are metaphysically queer. I offered the following definition of a non-natural property: a property, P, is non-natural if and only if instantiations of P are not token-token identical with the instantiations of any set of microphysical properties. Put another way, instantiations of non-natural properties are not token-token identical with any set of microphysical facts. Admittedly, this use of "non-natural" is idiosyncratic. In the remainder of this chapter it will be important to keep in mind that I am using "non-natural" in a way that is likely unfamiliar.

Predictive power and property identification

Over the last forty years, debates regarding the plausibility of reductionism have been prevalent in the philosophy of mind literature. The views I develop in this chapter are, like the views of many who approach metaphysics from a methodologically naturalist perspective (see, e.g, Ladyman et al.), deeply indebted to Daniel Dennett's work in the philosophy of mind. As befits the etiology of my thought, I will present my arguments against the backdrop of related debates in the philosophy of mind.

My primary meta-ethical opponent, the error theorist, holds that there are no moral properties. Error theory has an analog in the philosophy of mind: eliminative materialism. Just as the error theorist holds that there are no moral properties, the eliminative materialist holds that there are no mental properties.

In a variety of places, Daniel Dennett has argued against eliminative materialism. I rely heavily on the following thought experiment. I quote Dennett at length:

> Suppose … some beings of vastly superior intelligence—from Mars, let us say—were to descend upon us… They can be supposed to be Laplacean super-physicists, capable of comprehending the activity on Wall Street, for instance, at the microphysical level. Where we see brokers and buildings and sell orders and bids, they see vast congeries of subatomic particles milling about…
>
> Suppose … that one of the Martians were to engage in a predicting contest with an Earthling… From the Earthling's point of view, this is what is observed. The telephone rings in Mrs. Gardner's kitchen. She answers, and this is what she says: "Oh, hello dear. You're coming home early? Within the hour? And bringing the boss to dinner? Pick up a bottle of wine on the way home then, and drive carefully." On the basis of this observation, our Earthling predicts that a large metallic vehicle with rubber tires will come to a stop on the drive within one hour, disgorging two human beings, one of whom will be holding a paper bag containing a bottle containing an alcoholic fluid… The Martian makes the same prediction, but has to avail himself

> of much more information about an extraordinary number of interactions of which, so far as he can tell, the Earthling is entirely ignorant. For instance, the deceleration of the vehicle at intersection A, five miles from the house, without which there would have been a collision with another vehicle—whose collision course had to be laboriously calculated over some hundreds of meters by the Martian. The Earthling's performance would look like magic! (Dennett "True believers" 68-70)

Dennett is, without a doubt, correct that, from the perspective of the Martian, "the Earthling's performance would look like magic!" I take it that this vignette provides the seeds of good reason to be doubtful of eliminative materialism. One can find the clearest explanation of the reason the above vignette is a putative problem for the eliminative materialist in Dennett's *locus classicus*: "Real Patterns." How is the appearance of magic related to the reality of folk psychological states? The answer: it is all about predictive power. Dennett notes the following:

> We use folk psychology—interpretation of each other as believers, wanters, intenders, and the like—to predict what people will do next… Without its predictive power, we could have no interpersonal projects or relations at all; human activity would be just so much Brownian motion… *Where utter patternlessness or randomness prevails, nothing is predictable. The success of folk-psychological predictions, like the success of any prediction, depends on there being some order or pattern in the world to exploit.* [emphasis added] (Dennett "Real patterns" 30)

I take this to be one of Dennett's key insights. Where the comprehending application of some predicate provides predictive power, we have good reason to believe that the predicate tracks a genuine property. This point is easily illustrated.

Before doing so, however, it is important to include a brief aside regarding what I mean by "comprehending application." To help structure the following discussion, consider the following vignette. Imagine that you are at a zoo standing outside of an exhibit labeled "Zebras." Inside of the closure you see a black-and-white striped animal that looks, for all the world, like a zebra. Unbeknownst to you, the animal is a cleverly disguised mule (Dretske). Suppose you've named the animal Ed. You say, "Ed is zebra."

Ed is not a zebra. When you say, "Ed is a zebra" you have said something false. Something has gone wrong with your use of the predicate "is a zebra." There are, however, different ways someone's use of a predicate could go wrong. Imagine that, instead of saying "Ed is a zebra" you had said, "Ed is a starfish." Here, not only would you have said something false, it would further appear that you have failed to comprehend the concept *starfish*.[11] False statements can demonstrate one's facility with a concept. Thus, even though Ed is not a zebra, when you say, "Ed is a zebra," your utterance demonstrates your comprehension of the concept *zebra*.

The "comprehending application" of a predicate should be understood in terms of the speaker's facility with a concept. If a predicate is comprehendingly applied, the speaker has demonstrated her understanding of a concept. When a predicate is applied but not comprehendingly applied, the speaker has given one reason to doubt that the speaker understands the concept that corresponds with the predicate.

Comprehending application of a predicate is orthogonal to the truth of an utterance. Thus, when you said, "Ed is a zebra," you said something false; however, you also demonstrated your facility with the concept *zebra*. Just as one can say something false by comprehendingly applying a predicate, one can say something true while failing to comprehendingly apply a predicate. Imagine that, instead of saying "Ed is a zebra" you had said, "Ed is a mule." Here you have said something true. Ed is, in fact, a mule. However, you have not demonstrated your facility with the concept *mule*. If you had a complete grasp on the concept *mule*, you would not predicate, "is a mule" of creatures that look like Ed.

On this understanding of "comprehending application," at certain periods in history, one could have comprehendingly applied the predicate "is a witch" or "has

[11] This is, of course, assuming that the sentence "Ed is a starfish" is best interpreted as an attempt to express a belief. Language is often used to do something other than merely describe the world. If the utterance "Ed is a starfish" is not best interpreted as an attempt to describe the world, the utterance may show nothing about your understanding of the concept *starfish*.

phlogiston." Comprehending predication does not presuppose that a predicate corresponds with anything real.

Why should gains in predictive power as a consequence of the comprehending application of some predicate lead us to believe that the predicate tracks a genuine property (or properties)? Let "Fred" identify some existing animal. Imagine that, prior to introducing them to Fred, we tell subjects that the predicate "is an animal" has been comprehendingly applied to Fred. Subjects know nothing else about Fred. We then give them the following "test":

a. Fred has scales.

b. Fred has feathers.

c. Fred has hair. [12]

Presuming that Fred either has scales, feathers, or hair, e.g. that Fred is not a jellyfish, sea cucumber, or any such, one would expect subjects to perform exactly at chance.[13] Knowing nothing more about Fred than that he is an animal, subjects can do no better than guess about Fred's exterior attributes. But now imagine that we tell test subjects that the predicate "is a bird" has been comprehendingly applied to Fred. One can now expect near 100% accuracy on the test. Just about everyone will correctly answer (b): Fred has feathers. Note that this change in predictive accuracy is (1) publically

[12] If you're worried about artificially restricting the possible types of animal that Fred could be, feel free to drop this restriction and fill out the test with the necessary descriptions of possible dermal layers. So long as the test offers finite and mutually incompatible answers, responses should be at chance.

[13] "At chance" will be determined by the distribution of animals with scales, feathers, and hair. If there are an equal number of animals with scales, feathers, and hair, subjects will answer correctly approximately 33% of the time. If, of the animals that have either scales, feathers, or hair, 50% have scales, 30% have feathers, and 20% have hair, subjects who guess (a) will answer correctly 50% of the time, subjects who answer (b) will answer correctly 30% of the time, and subjects who answer (c) will answer correctly 20% of the time.

observable and (2) repeatable. As such, it constitutes a scientifically acceptable phenomenon. One can now demand an explanation and the scientific realist is committed to taking the best explanation of this phenomenon to be indicative of ontology.

Before proceeding, I should say a bit more about the kind of test I have in mind. In every instance, a predicate is comprehendingly applied to an object. The object must be an existent. Test subjects are then told that a predicate was comprehendingly applied to an object and are asked to select, from a list of options, the statement that is mostly likely to be true, given what they know about the subject. These options are all statements that can be observationally confirmed, where we understand "observational confirmation" is an appropriately theory laden way. Given that truth and comprehending application can come apart, little should be read from a single instance of applying a predicate to a subject. Rather, we will need to run repeated tests, changing the subject each time. Over the course of repeated tests, for most predicates, we can expect to see consistent improvement above chance in test subject's ability to select a true statement from the available options (for at least some sets of questions). What does this tell us?

Consider our original example: the comprehending application of the predicate "is a bird" to Fred. Suppose that the predicate "is a bird" did not correspond with a genuine property. Would we expect predictive power to improve? Let "is a schmird" be a predicate that fails to correspond to any property. Further suppose that test subjects are told that the predicate "is a schmird" was accurately applied to Fred. Controlling for framing effects, one would expect no increase in predictive accuracy. In virtue of not corresponding to any property, comprehending application of "is a schmird" cannot tell test subjects anything about the status of Fred's dermal layer. Analogously, if "is a bird" did not track some property or set of properties, one would not expect any improvement in subjects' predictive powers. If a predicate does not correspond to a property or set of properties, application of the predicate tells subjects *nothing* about Fred, leaving them in no better position to make an accurate prediction than they were before they learned that *Fred is a schmird.*

This is why the magical appearance of the Earthling's predictive powers is relevant to ontology: "the success of any prediction… depends on there being some order or pattern in the world to exploit" (Dennett "Real patterns" 30). There is one, and only one, explanation of the predictive improvement offered by the comprehending application of a predicate. The predicate *must* co-vary with some property or properties. Anything else leaves the improvement in predictive power unexplained: a miracle.

The ethical non-naturalist needs a way to demonstrate, contra the error theorist, that there are moral properties. If the ethical non-naturalist can show that the comprehending application of moral predicates improves predictive power, the ethical non-naturalist will have given sufficient reason for believing that the comprehending application of moral predicates co-varies with the presence of some set of properties. Put another way, if the comprehending application of moral predicates improves predictive power, we have good reason to believe that whenever we comprehendingly apply a moral predicate some particular property or set of properties has been instantiated.

It is, however, important not to overstate the significance of this conclusion. If one can show that comprehending application of a predicate improves one's predictive power, one has shown that the relevant predicate co-varies with *some* property or properties. This is not, however, to say anything about the property that is being tracked. Earlier I noted that, as I was using the phrase, the predicates "phlogiston" and "witch" could be comprehendingly applied. Note that comprehending application of either predicate appears to allow one to make predictions at better-than-chance odds. Suppose "witch" was comprehendingly applied to some person, P. We could have predicted that

P was elderly, a female, and impoverished[14] *and* we could have made these predictions with greater accuracy than would be possible by merely matching our guesses with sociological data, e.g. predicting that "P was female" 50% of the time just in case 50% of the population was female.

Let us suppose the best-case scenario for the ethical non-naturalist: comprehending application of moral predicates allows us to make some predictions with better-than-chance accuracy. Were this the case, it would be a boon for ethical non-naturalism. The ethical non-naturalist could then make the following claim: we have good reason to believe that our moral predicates co-vary with some set of properties. A lot more, however, needs to be said in defense of ethical non-naturalism. The ethical non-naturalist will say the following: moral predicates co-vary with non-natural moral properties. The error theorist will say something entirely different: just as the comprehending application of the predicate "is a witch" co-varies with non-witch properties, moral predicates co-vary with non-moral properties.

Explanations and mind-independent properties

I have now offered a method for identifying a set of properties: if comprehending application of a predicate improves predictive power, we have good reason to believe that comprehending application of the predicate co-varies with the instantiation of some particular set of properties. The remainder of this chapter will be dedicated to exploring various ways we might go about characterizing the properties in question. By the end of the chapter, I hope to have compiled a list of criteria such that, if a predicate meets all of the criteria, we have good reason to believe that comprehending

[14] Given that not every elderly, impoverished female was considered a witch, presumably fulfilling these three conditions was not sufficient for the predicate "is a witch" to be comprehendingly applied to someone.

application of the predicate co-varies with the instantiation of a mind-independent non-natural property.

Importantly, at this juncture, the debate between the ethical non-naturalist and the error theorist is a debate about best explanations. Some explanation is owed of the predictive power we gain by applying various predicates. The first step in providing such an explanation is to note that the predicate in question must be tracking some property or some set of properties; however, at best, this observation provides a starting point for an explanation. Some further characterization of the properties in question is owed and this further characterization must have the resources to provide the explanatory tools necessary to link the properties in question with the gains in predictive power we see. The key for the ethical non-naturalist will be to find a set of conditions such that, if a predicate meets these conditions, then the best explanation of the predictive power provided by the application of the predicate is that comprehending application of the predicate co-varies with mind-independent non-natural properties.

The error theorist is committed to the falsity, not just of ethical non-naturalism, but of moral realism more generally. As I have characterized moral realism, the realist is committed to the mind-independence of moral properties. This suggests the following explanatory move on behalf of the error theorist: the predictive gains (if any) we see in virtue of comprehendingly applying normative predicates can be explained by co-variance between the comprehending application of normative predicates and people's mental states. If this explanatory strategy is successful, showing that the comprehending application of normative predicates improves predictive power will offer no succor to the moral realist. The only properties that have been identified by this method are mind-dependent and, by the realist's hypothesis, mind-dependent properties cannot be the moral properties.

An example might help to illustrate. Suppose we comprehendingly predicate "is good" of some property, e.g. "beauty is good." Further suppose that this predication allows us to predict that Holly will act in ways that increase the amount of beauty in the world. The error theorist can plausibly explain this kind of improvement in predictive power by holding that the predicate "is good" tracks sentiments of approval or disapproval. Why does believing that "beauty is good" allow us to comprehendingly predict that Holly will act in ways that increase the amount of beauty in the world? Because our application of the predicate "is good" reliably co-varies with Holly's approval.

I have no concrete suggestions of what criteria the non-naturalist could propose that might rule out the success of this explanatory strategy. Explanations must be considered on a case-by-case basis. It is clear that the predictive power provided by the application of some predicates cannot plausibly be explained by claiming that the predicate in question reliably co-varies with anyone's mental states. Consider the original example I offered. If we comprehendingly apply the predicate "is a bird" to Fred, subjects will successfully predict that Fred has feathers. No explanation, in terms of mental starts, of the improvement in predictive power offered by the comprehending application of the predicate will be forthcoming. Facts about the covering of Fred's epidermis are independent of any of our mental states. Consequently, we will be unable to explain the predictive power gained by application of the predicate "is a bird" in terms of co-variance with some set of mental states. A successful explanation will require that the predicate co-vary with some set of external world properties.

We can make the point somewhat more general. Suppose that the comprehending application of some predicate improved predictive power. The best explanation of this improvement in predictive power is that comprehending application of the predicate co-varies with a property. How does this putative co-variance explain why comprehending application of the predicate improves predictive power? The comprehending application of a predicate will allow us to make accurate predictions just in case comprehending application of the predicate co-varies with some property, P, and

P co-varies with the property about which we are making a prediction. The explanation works by showing that there is a relationship between comprehending application of a predicate and the state-of-affairs about which one is making a prediction—the relationship exists in virtue of the predicate co-varying with some property, P, and P co-varying with the predicted state-of-affairs. An example ought to help. Applying the predicate "is a bird" to Fred allows us to predict that Fred is a bird because the use of the predicate "is a bird" co-varies with the property of *being a bird* and the property of *being a bird* is nomologically linked to the property of *having feathers*.

In cases where co-variance with mental states is going to explain how the application of a predicate can improve our predictive accuracy, there must be some explanatory relationship between mental states and the predicted property. When no such explanatory relationship exists, mental states cannot explain the predictive power gained by the comprehending application of a predicate. Our best scientific theories tell us when it is plausible to think that an explanatory relationship holds between some set of mental states and the accuracy of some prediction. A case-by-case examination of instances in which predicates improve predictive power is likely necessary to determine when an explanation in terms of predicate co-variance with mental states is available.

We now have two criteria a predicate needs to meet in order to help the ethical non-naturalist make her case. First, comprehending application of a predicate must improve predictive power. Second, the best explanation of the previous is not in terms of a predicate co-varying with mental states.

Identifying the absence of token-token identities

If the predictive power gained by the comprehending application of a predicate cannot be explained by the co-variance of a predicate and some set of mental states, we have good reason to believe that a predicate co-varies with some set of mind-independent properties. That is, we have good reason to believe that the properties in questions are candidates for moral properties, realistically construed.

I have argued that the ethical non-naturalist is committed to the existence of moral facts that are not token-token identical with microphysical facts. The non-naturalist is committed to the claim that the moral facts are not token-token identical with any set of microphysical facts. Given my strategy for defending ethical non-naturalism, this amounts to the ethical non-naturalist being committed to the claim that *the facts that co-vary with normative predicates are not token-token identical with microphysical facts*. The ethical non-naturalist owes us some account of what would have to be true about a predicate for it to be the case that the facts with which the predicate co-varies are not token-token identical with microphysical facts. In the remainder of this section, I will develop a criterion that enables us to test for token-token identities with microphysical facts.

Abstraction

Dennett has famously—or perhaps more accurately, infamously—defended a heuristic account of intentionality: the Intentional stance. The Intentional stance is exceedingly hard to characterize, attempting to occupy an uncomfortable middle ground between realism and anti-realism regarding intentionality. Nonetheless, Dennett's thought on the subject is instructive. Again, I quote him at length:

> Consider the case of a chess-playing computer, and the different strategies or stances one might adopt as its opponent in trying to predict its moves. There are three difference stances of interest to us. First there is the *design stance*. If one knows exactly how the computer is designed… one can predict its designed response to any move one makes by following the computation instructions of the program. One's prediction will come true provided only that the computer performs as designed—that is, without breakdown… [O]ne can make design-stance predictions of the computer's response at several different levels of abstraction, depending on whether one's design treats as smallest functional elements strategy-generators and consequence-testers, multipliers and dividers, or transistors and switches…

> Second, there is what we may call the *physical stance*. From this stance our predictions are based on the actual physical state of the particular object, and are worked out by applying whatever knowledge we have of the laws of nature. It is from this stance alone that we can predict the malfunction of systems…
>
> The best chess-playing computer these days are practically inaccessible to prediction from either the design stance or the physical stance; they have become too complex for even their own designers to view from the design stance. A man's [sic] best hope of defeating such a machine in a chess match is to predict its responses by figuring out as best he can what the best or most rational move would be, given the rules and goals of chess… Put another way, when one can no longer hope to beat the machine by utilizing one's knowledge of physics or programming to anticipate its responses, one may still be able to avoid defeat by [taking the Intentional stance and] treating the machine rather like an intelligent human opponent. (Dennett "Intentional systems" 87-89)

For many years, the philosophy of mind was under the computationalist spell: the reigning paradigm of research into the mind took the mind to be closely analogous to a computer. In his description of the three stances one can use as prediction heuristics, Dennett self-consciously models a prevalent theme in computer science: abstraction.

Colburn and Shute take abstraction in computer science to consist in "hiding" information. They nicely illustrate the centrality of abstraction for computer science:

> A computational process for an electronic digital computer is described statically, and at varying levels of detail, in textual artifacts written in programming languages. Depending on the type of programming language, whether it be machine language, assembly language, or any of a number of kinds of higher-level languages, the elements of computational processes that are described in textual programs might be as basic as binary digits (bits) and machine registers, or as familiar as telephone books and shopping carts. Whatever the elements of computational processes that are described in textual programs, however, they are never the actual, micron-level electronic events of the executing program; textual programs are always, no matter what their level, *abstractions* of the electronic events that will ultimately occur. Since it is a practical impossibility for a textual program to describe the electronic events, every such program describes a computational process as an abstraction that hides information or details at a lower level. (Colburn and Shute 177)

One might think that everything there is to be said about a particular computer running a particular piece of software can be said in the language of the physical instantiation of the machine. One need not look any further than a purely mechanical description to capture absolutely everything that occurs in a computer. But this is not a particularly helpful way of proceeding. Programming in assembly code takes less time than programming in machine code. Either takes far less time than attempting to conceptualize one's computer programs in terms of the mechanical instantiations of logic gates. By "hiding" information, the computer scientist makes the job of writing software significantly easier.

Dennett's three stances are nothing more than three levels of abstraction dressed in their Sunday best. When one adopts the physical stance, one approaches a system without any abstraction. No information is hidden; no detail is left out. When one adopts the design stance, one approaches a system from an intermediate level of abstraction. All of the details of a program's physical instantiation are hidden. Programs are approached in terms of their *functional roles*, leaving aside any question of how these functional roles are instantiated. Lastly, when one adopts the Intentional stance, one approaches a system from a high level of abstraction. Information about a program's physical instantiation *and* information about a program's design are both obscured.

Abstraction is a linguistic phenomenon. It is a question of the way we conceptualize systems—of our choice of language. We can describe things in more detail, e.g. take up the physical stance, or we can describe things in less detail, e.g. take up the intentional stance. Abstractions hide information that is available about one and the same phenomenon described at a lower level of abstraction.

I'd like to formalize what I have, so far, said about abstraction. Before doing so, however, I must introduce a new technical term: *basic sufficient instantiater*. For any true proposition about a token instance of a property instantiation, the *basic sufficient instantiater* is the most ontologically basic obtaining state-of-affairs[15] sufficient for the truth of the proposition.[16] An example should help illustrate.

Consider a mousetrap. Call this mousetrap, M. Call the following proposition[17] "[M]": *M is a mousetrap.* M's having the property of *being designed to trap mice* is sufficient for the truth of [M]. However, *M's having the property of being designed to trap mice* may not be the most ontologically basic obtaining state-of-affairs which is sufficient for [M]'s being true. If M's being a mousetrap is token-token identical to a set of microphysical facts, the state-of-affairs constituted by the relevant set of microphysical facts is [M]'s basic sufficient instantiater. These microphysical facts would constitute the most ontologically basic obtaining state-of-affairs sufficient for the truth of [M].[18]

Let p_1 and p_2 denote two distinct predicates. Under what conditions would we want to say that a sentence including p_1 is an abstraction of a sentence including p_2? For ease of reference, let P_1 denote some sentence where p_1 is predicated of an entity, e.g. "x is p_1," and let P_2 denote some sentence where p_2 is predicated of the same entity, e.g. "x is p_2." We can then reformulate the question: under what conditions is P_1 an abstraction

[15] I wish to remain agnostic about the exact nature of states-of-affairs. If your preferred account of states-of-affairs renders my account of sufficient basic instantiaters problematic, feel free to substitute terminology as necessary.

[16] In those instances where a *sui generis* property supervenes on some subvenient base, for the purposes of determining the basic sufficient instantiater, always consider the subvenient base more ontologically basic than the supervening property.

[17] I am, here, using "proposition" as a placeholder. Feel free to insert whatever one takes to be the bearer of truth.

[18] It should be noted that multiple realizability is orthogonal to a property instantiation's basic sufficient instantiater. Only propositions about token instantiations of properties have basic sufficient instantiaters. Let P be some proposition about a token instantiation of a property and let T_1 be the subject of P. One cannot change the basic sufficient instantiater of T_1 without thereby changing the subject of P, i.e. without thereby introducing an entirely new proposition. Put another way, P's basic sufficient instantiater is an identity condition of T_1.

of P_2? For the time being, I will offer a partial answer: if P_1 is an abstraction of P_2, then P_1 and P_2 have the same basic sufficient instantiater.

The point is perhaps made most obviously via example. Label a particular patch of imperial red "R." Presume that a secondary-property analysis of *redness* is accurate.[19] Consider the following abstraction: R is colored. What is the basic sufficient instantiater of the proposition? The answer seems to be: the facts referred to by the lowest level abstracted description of R, i.e. the facts picked out by the claim that R *has such-and-such surface reflectance properties.* Consider a level of abstraction previous to "R is colored": R *is red.* If R is red, then R is colored. Furthermore, if R is imperial red, then R is red. Thus, the truth of the proposition R *is imperial red* is sufficient for the truth of the proposition R *is colored.* Consequently, the obtaining of the state-of-affairs picked out by the proposition R *is imperial red* is sufficient for the truth of the proposition R *is colored.*

We can put the point schematically. Consider some claim about R made at an arbitrary level of abstraction, A_n. The truth of the claim about R made at the A_n level of abstraction is a necessary condition for the truth of a claim made at a previous level of abstraction, A_{n-1}. The same goes for the truth of the claim about R made at A_{n-1}. The truth of the claim about R at the A_{n-1} level is a necessary condition for the truth of a claim made at a previous level of abstraction, A_{n-2}. But if the truth of a claim about R made at A_n is a necessary condition for the truth of a claim made about R at A_{n-1}, and the truth of a claim about R made at A_{n-1} is a necessary condition for the truth of a claim made about R at A_{n-2}, then the truth of a claim about R made at A_n is a necessary condition for the truth of a claim about R at A_{n-2}. *It follows that the truth of a claim about* R *made at A_{n-2} is sufficient for the truth of any other claim about* R *made at any higher level of*

[19] For the example to work, we must also hold constant a range of other facts, e.g. facts about what constitutes "normal lighting" and a "normal observer."

abstraction. Consequently, all claims about R made at some level of abstraction have the same basic sufficient instantiater. If P_1 is an abstraction of P_2, then P_1 has the same basic sufficient instantiater as P_2.

Sharing a basic sufficient instantiater is not enough to make P_1 an abstraction of P_2. The following two propositions have the same basic sufficient instantiater but neither is an abstraction of the other: "Clark Kent has black hair" and "Superman has black hair." There is a further condition that must be met. If P_1 is an abstraction of P_2, then P_1 must hide implementation details that P_2 reveals *and* P_1 must not reveal implementation details that P_2 hides. This nets us the following conditions that P_1 must meet in order to be an abstraction of P_2: (1) P_1 and P_2 must have the same basic sufficient instantiater and (2) P_1 must hide implementation details that P_2 reveals (while not revealing implementation details that P_2 hides).[20]

While fulfilling the above two conditions are constitutive of P_1's being an abstraction of P_2, if P_1 is an abstraction of P_2, there is a further relationship that we can expect to hold between P_1 and P_2. If P_1 is an abstraction of P_2, we can expect predictions based on P_1 to be less accurate than predictions based on P_2; however, we can also expect predictions based on P_1 to require less computational power than predictions based on P_2.

[20] In the remainder of the dissertation I will talk about *sentences*, *propositions*, and *predicates* being abstractions. For the purposes of the dissertation, there is no need for me to take a stance on the exact relationship between sentences and propositions. Consequently, I will use the two terms interchangeably. Since we can make sense of talk about basic sufficient instantiaters regarding both sentences and propositions, there is nothing *prima facie* problematic about treating both sentences and propositions as abstractions. The same cannot be said about predicates. Stand-alone predicates are not descriptions of the world and, consequently, cannot be either true or false. Nonetheless, I think that, for reasons of clarity, it can be helpful to talk about predicates as if they were abstractions. All such locutions should be taken to be enthymematic for some range of sentences in which the predicate in question is predicated of some particular set of subjects.

Before making this case, I need to say a bit more about what I mean by both "computational power" and "predictive accuracy." As I am using the term, "computational power" can be understood in terms of operations performed over bits. A bit is a measure of information storage. We can think of bits as switches. Every switch is either *on* or *off*. We can use the position of a switch, either *on* or *off*, to store information. Operations can then be understood in terms of flipping a switch. So if we flip a single switch from *on* to *off*, we have performed a single operation on a single bit. Computational power can be measured in terms of operations. A calculation that requires five operations requires more computational power than a calculation that requires three operations.

So far, I have been approaching predictive accuracy in terms of the probability of getting the correct answer. We can helpfully think about this notion of predictive accuracy in terms of a multiple-choice test. If a question on such a test has three answers, without even reading the question, we can expect to get the right answer 33% of the time. Any increase in this probability counts as an improvement in predictive accuracy.

Consider again the wager Dennett imagines between a Laplacean Martian and an Earthling. In order to make its prediction, the Martian made an extraordinary number of calculations. By contrast, the Earthling's prediction took staggeringly little computational power. Why the difference? The Martian used the physical stance to make its prediction. The Earthling used the Intentional stance to make hers. Abstraction allows one to make predictions while using comparatively small amounts of computational power. One need not take into account all of the minutia of physical instantiation, or even the minutia of design details, in order to make one's predictions. Thus, the Earthling can make predictions that, to the Martian, appear to come at a miraculously small price in computational power.

Put another way, for the Martian to successfully make its prediction, the Martian had to assign a bit to every sub-atomic particle. The Martian's calculation then required that the Martian run multiple operations over every such bit. By embracing the Intentional stance, the Earthling massively reduced the number of entities for which she must assign bits, and correspondingly, massively reduced the number of operations required to make her prediction.

Of course, a reduction in the computational power required to make a prediction does not come for free. That would, indeed, be a miracle. Instead, there is a trade-off. Decreased need for computational power comes at the price of predictive accuracy. Consider again our Laplacean Martian. In virtue of taking the physical stance, the Martian is in a position to take such things as car accidents, flat tires, and freak tornadoes into account. The Earthling, using the Intentional stance to make her predictions, does not have the resources to factor any of these contingencies into her prediction. The Martin's predictions come at a staggering cost in terms of computational power; however, they are one hundred percent accurate. The Earthling's predictions are comparatively cheap; however, they are also much less accurate. This trade-off is not unique to the Martian and the Earthling. Every time one moves up a level in abstraction, e.g. from the physical stance to the design stance, one's calculations need take into account fewer variables, so predictions come cheaper; however, one is no longer in a position to take into account the idiosyncrasies of the previous level of abstraction. As a result, as one climbs the levels of abstraction, the accuracy of one's predictions decreases.[21]

[21] A further word is owed about the above trade-offs. One cannot make any prediction based on a single abstraction. Successful prediction will require some set of auxiliary hypotheses. A restriction needs to be made regarding the auxiliary hypotheses that one gets to draw on in making predictions. We need to rule out any set of auxiliary hypotheses that allows one to reconstruct the implementation details hidden by the current level of abstraction. Consider the following abstraction: "R is colored." The above trade-offs fail to hold if we are allowed to draw

This trade-off, between computational power and predictive accuracy, allows us to use language, in the form of abstraction, as a test for token-token identities. Consider two propositions, P_1 and P_2. P_1 will be an abstraction of P_2 if, and only if, P_1 has the same basic sufficient instantiater as P_2 *and* P_1 hides information that P_2 makes available, i.e. P_1 is at a higher level of generality than is P_2. This is just what it is to be an abstraction. So if P_1 fails to be an abstraction of P_2, it follows that either (1) P_1 does not have the same basic sufficient instantiater as P_2 or (2) P_1 does not hide information that P_2 makes available, i.e. P_2 is at the same, or a higher, level of generality as P_1. Given two propositions, P_1 and P_2, if we have good reason to believe that P_1 is at a higher level of generality than P_2 and that P_1 is not an abstraction of P_2, we have good reason to believe that P_1 and P_2 do not share a basic sufficient instantiater. The abstraction relationship is characterized by a trade-off between predictive accuracy and computational power. Where this relationship does not hold between P_1 and P_2, we have good reason to believe that the state-of-affairs picked out by P_1 is not token-token identical to the state-of-affairs picked out by P_2.

Abstraction and meta-ethics

I have now sketched a rough strategy for identifying cases in which two sets of facts are not token-token identical. If the abstraction relationship does not hold between two propositions that predicate different properties of the same entity then, presuming we have good reason to suppose that one proposition is at a higher level of generality

on an auxiliary hypothesis of the form: "If R is colored then R is imperial red." In virtue of revealing the implementation details that were otherwise hidden by the abstraction, these kind of auxiliary hypotheses functionally remove the status of abstraction from a given predicate.

This condition on acceptable auxiliary hypotheses may appear *ad hoc*. It is not. The aim of the current section is to establish a test for abstraction. It is no criticism of a test that, in order to be effective, we have to specify the conditions under which it is run.

than the other, we have good reason to believe that the two properties are not token-token identical. Having introduced the notion of abstraction and the way in which it can ground a test for token-token identities, it is time to relate these notions back to meta-ethics.

I have two primary opponents: the reductive realist and the error theorist. The abstraction test is not straightforwardly relevant to the error theorist's thesis.[22] The trouble is that the error theorist does not think that moral propositions are ever true. That is, the error theorist is already convinced that the properties referred to by normative predicates are not token-token identical with microphysical properties. Normative predicates *fail to refer at all*. This forces us to search for a thesis about abstraction that the error theorist must accept.

Connecting error theory to the abstraction test will be somewhat convoluted. I'll start by introducing two new classes of predicates. Let each member of the set {m_1, m_2, m_3, … m_n} represent some proposition that (1) predicates of some entity a normative property, e.g. "x is right," and (2) is such that acceptance of the proposition improves our ability to make a certain class of predictions and it is implausible to suppose that this improvement in predictive power is a consequence of the predicate co-varying with mind-dependent properties. For each member of the set {m_1, m_2, m_3, … m_n} there will be some set of external world properties that co-vary with the proposition in virtue of which acceptance of the proposition improves our ability to make a certain class of

[22] The intuitive approach is to take some sentence including a normative predicate, e.g. "x is wrong," and show that it is not an abstraction of some complicated claim about microphysical properties. I've suggested that this would show that the proposition "x is wrong" does not have the same basic sufficient instantiater as claims about microphysical properties and thereby show that the properties that co-vary with the comprehending application of normative predicates are not token-token identical with microphysical properties. This strategy can be successful against a certain brand of reductive realist, i.e. the reductive realist that is committed to thinking that "x is wrong" is true *and* is made true by a complicated set of microphysical facts. The strategy will not, however, be successful against the error theorist.

predictions. Let each member of the set $\{n_1, n_2, n_3, \ldots n_n\}$ be paired with a member of the set $\{m_1, m_2, m_3, \ldots m_n\}$; thus, n_x is paired with m_x. Let each n_x be a definite description of the form "the external world properties tracked by m_x in virtue of which acceptance of m_x improves our ability to make predictions." Call each member of the set $\{n_1, n_2, n_3, \ldots n_n\}$ an "N-predicate" and the set taken as a whole the class of "N-predicates." In virtue of N-predicates consisting of definite descriptions, the referent of N-predicates is opaque. Each N-predicate picks out some set—*we know not which*—of external world properties.

Let me introduce a further set of predicates. N-predicates pick out some set of external world properties via opaque definite description. Let "P-predicates" correspond one-to-one with N-predicates. P-predicates pick out all and only the same properties as N-predicates; however, instead of consisting of the definite description "*whatever* properties it is in virtue of which acceptance of m_x improves our predictive capabilities," let P-predicates be a description of the properties responsible for the relevant improvement in predictive power. The ethical non-naturalist will think that P-predicates are constituted by descriptions of non-natural normative properties. The error theorist will think that P-predicates are composed of some complicated description of sets of microphysical properties. We are now in a position to relate abstractions to the error theorist's thesis.

The error theorist is committed to thinking that sentences including N-predicates are abstractions of sentences including P-predicates; the error theorist is committed to thinking that the properties tracked in virtue of which the comprehending application of normative predicates can improve predictive accuracy are microphysical properties. Suppose that one thinks that the relevant properties are some kind of emergent property, i.e. not microphysical properties. The non-naturalist can happily accept token-token identities between normative facts and non-microphysical facts. If the error theorist wants, at this point in the argument, to resist the non-naturalist's thesis, she must think that P-predicates are composed of some complicated description of microphysical facts. It would follow that the state-of-affairs picked out by P-predicates are the basic sufficient instantiater or sentences including N-predicates.

N-predicates reveal *no* implementation details, whereas P-predicates are in the business of revealing implementation details. It follows that the error theorist is committed to thinking that sentences including N-predicates are abstractions of sentences including P-predicates. If we can show that N-predicates are not abstractions of P-predicates, we will have shown that the facts that co-vary with the comprehending application of normative predicates are not token-token identical to microphysical facts.

NP-hard problems and the failure of token-token identities

The following section may be a bit difficult to follow—a brief summary is in order. I previously argued that the abstraction relationship is characterized by a trade-off between predictive accuracy and computational power. If we can show that this tradeoff does not hold between N-predicates and P-predicates, we can show that N-predicates are not abstractions of P-predicates. There are cosmological limits on the computational power of the universe. That is, if the universe were one enormous computer, over the course of its entire lifetime, it would only have so much computational power. Suppose that we can make a prediction, Z, at above chance based on the comprehending application of an N-predicate. Further suppose that predicting Z based on the comprehending application of a P-predicate would have to be at chance because there is not enough computational power in the universe to predict Z based on the properties that reliably co-vary with the comprehending application of the P-predicate. It would follow that the N-predicate in question is not an abstraction of a P-predicate. Were the N-predicate an abstraction of a P-predicate, we would expect predictions based on the N-predicate to be less accurate than predictions based on the P-predicate. But in the imagined situation, predictions based on the N-predicate are *more* accurate than predictions based on the P-predicate. By the conclusion of this section I will have shown, via consideration of the cosmological limits on computational power, that any N-predicate the comprehending application of which improves predictive power at or above the organism level is not an abstraction of a P-predicate.

We will, once again, have to make something of a brief detour, this time through computational complexity theory. I quote Unger and Moult at length:

> The notion of NP-completeness was invented to describe a class of problems that are "hard" to solve. For all of the problems in the class there exists an *exponential* time algorithm, but a *polynomial* time algorithm is not available for any of them. Consider for example the Hamiltonian path problem: Given a graph (e.g. a road map between *n* cities), is there a path that visits each node (e.g. city) exactly once? It is clear that by examining all of the exponentially many possible paths one can decide whether a Hamiltonian path exists, but a polynomial time algorithm is not known. The terms exponential and polynomial measure the type of dependence of the running time of the algorithm (until a solution is found) on the size (based on a reasonable representation) of the data. The dependence is based on a "worst case" analysis, namely the time that guarantees to bound the performance of the algorithm on every possible instance of the data. An algorithm is said to be polynomial if the running time can be bounded by a polynomial function in the size *n* of the problem, and exponential if the dependence is described by an exponential function of *n*. The distinction between exponential time and polynomial time solutions is crucial because of a very simple fact: exponential functions grow much faster that polynomial ones. While an algorithm that has a polynomial running time (even if the polynomial function is of a relatively high order) is feasible on modern computers even for big problems, exponential algorithms are useful only for very small "toy models". For example, if the size of the problem, *n*, is 100 an n^5 polynomial algorithm will take 10^{10} time units, which is feasible, but a 2^n exponential algorithm will require about 10^{30} time units, which is prohibitively long. If, for example, the time unit is a microsecond then the polynomial algorithm will take less than 3 hr while the exponential algorithm will require about 10^{16} years(!). (Unger and Moult 1185-1186)

Finding the solution to an NP-complete problem of any significant size is infeasible. NP-hard problems are the next more difficult set of problems; NP-hard problems are *at least* as difficult as the hardest NP-complete problems (Unger and Moult 1187).

Proteins are composed of amino acids. Scientists talk about three distinct and increasingly complex ways of thinking about protein structure. One can talk about a protein's primary structure, it's secondary structure, and its tertiary structure.[23]A protein's primary structure is just the list of the amino acids of which it is composed. A protein's secondary structure consists of three-dimensional structural motifs, e.g. helices that are formed as a consequence of the interactions between the amino acids of which the protein is composed. Finally, the tertiary structure of a protein is the three dimensional structure of the entire protein.

Unger and Moult have proven that determining the tertiary structure of proteins based on their primary structure is an NP-Hard problem (Unger and Moult). Consequently, we have some reason to believe that calculating the tertiary structure of a protein based on the amino acids of which it is composed will require 2^n operations, where n is equal to the number of amino acids of which the protein is composed. Calculating the tertiary structure of a protein composed of ten amino acids would require 2^{10} operations. Calculating the tertiary structure of a protein composed of twenty amino acids would require 2^{20} operations.

Presuming that calculating the tertiary structure of a protein based on its primary structure is the only NP-Hard problem in play, how long would it take to predict the tertiary structure of a single protein based on the properties of its component parts? Depending on the protein in question, the answer appears to be: forever. Computations are physically instantiated. The length of time it would take to perform some computation is dependent on the speed and power of the computer that one has running one's computation. But, as Krauss and Starkman write: "The physical nature of

[23] One can also talk about a protein's quaternary structure. For my purposes, the difference between tertiary and quaternary structure are unimportant—I will treat them as if they are the same and use the single label "tertiary structure" to pick out both.

computation… implies fundamental limits on the amount of information processing that finite physical systems can perform" (Krauss and Starkman 1). Tipler elaborates:

> [T]he ultimate physical limitations on computation will likely arise from limitations on the bit size of computer memories and the speed with which different parts of the computer can communicate with each other. The ultimate limit to physical size is the size of the entire universe, and the ultimate speed is that imposed by relativity. Thus, the ultimate physical limitations are those imposed by cosmology and relativity. (617)

The upshot is that "[there is] an ultimate limit on the processing capability of any system in the future, independent of its physical manifestation" (Krauss and Starkman 3). Imagine that the entire universe—every constitutive particle—composed one enormous computer. We can then ask this question about the computer: over the course of its lifetime, how many operations could the computer run? Modern physics allows us to make some predictions.

From start to finish, computation will be possible in the universe for 10^{100} years (Adams 104). In these 10^{100} years, how many operations could be performed? We'll need a bit of basic math and a few other estimates to find an answer. The predicted current age of the universe is 10^{10} years (Lloyd 3). In these 10^{10} years, "the universe can have performed no more than 10^{120}" operations (Lloyd 1). If the universe can perform 10^{120} operations in 10^{10} years, how many operations can it perform in 10^{100} years? The universe can perform ($10^{120}/10^{10}$) operations in one year. That is, the universe can perform 10^{110} operations a year. It follows that it can perform (10^{110} x 10^{100}) operations in 10^{100} years. So, were the universe one huge computer, in the course of its entire lifespan, it could run approximately 10^{210} operations. What does this tell us about computing the tertiary structure of proteins?

A single titin protein, found in muscles, is composed of "nearly 27,000 amino acids" (Wrigley 738). Calculation of the tertiary structure of a *single* titin protein is a computational task of astounding complexity. Given that such a calculation is NP-hard, one can expect such a calculation to take, at minimum, 2^{27000} operations. How does this compare to the total number of operations the universe could possibly compute, over its entire lifespan? We can display 2^{27000} as [(2^4)^6750]. Note that $2^4 > 10$. For simplification purposes, let us pretend that $2^4 = 10$, which allows us to represent "[(2^4)^6750)]" as "10^{6750}," all the while knowing that the actual number of operations it would take to compute the tertiary structure of a titin protein is greater than 10^{6750}. The computational upshot should be clear. On our best estimates, the universe could, over its entire lifetime, run 10^{210} operations. The calculation of the tertiary structure of a *single* titin protein requires (approximately) 10^{6750} operations. Even were the entire universe a single computer that spent its entire lifespan attempting to calculate the tertiary structure of a titin protein based on the titin's primary structure, the calculation would never be complete. This will remain the case even if we *substantially* revise our opinions about the number of operations the universe can carry out a year, or the number of years the universe would be capable of carrying out computations. The number of operations required to calculate the tertiary structure of a titin protein is 6540 *orders of magnitude higher* than the predicted number of operations the entire universe could possibly run. [24]

What does all of this have to tell us about abstractions? Remember that abstractions are characterized by a tradeoff between computational power and predictive accuracy. Moving up levels in abstraction decreases the amount of computational power required to make a prediction; however, it also decreases the accuracy of one's predictions. As the amount of computational power saved goes to infinity, the improvement above chance one's predictions gain by acceptance of the abstraction falls

[24] This says nothing about calculating the behavior of amino acids based on the properties of the constituents of amino acids, e.g. atoms, protons, neutrons, electrons, quarks, etc. Furthermore, it says nothing about calculating the interactions of a single titin protein with the rest of the constituents of a muscle, or calculating how a single muscle would interact with the rest of the human body. In terms of predicting the behavior of a single human based on the composition of the human body, calculating the tertiary structure of a titin protein is a tiny, though already impossible, snapshot of all of the calculations that would be required.

to zero. But if there are limits on the amount of computational power in the universe, we do not need to drive the amount of computational power saved to infinity in order to drop expected predictive gains to zero.

Consider two propositions, P_1 and P_2. Suppose that P_1 is at a lower level of generality than P_2, e.g. P_1 is a description of a cell at the microphysical level whereas P_2 is a functional description of a cell. Further suppose that one can make some prediction, R, on the basis of the comprehending application of P_2 that one cannot make based on comprehending application of P_1—because the calculation is nomologically impossible to complete. Under such conditions we have very good reason to believe that P_2 *is not* an abstraction of P_1. Were P_2 an abstraction of P_1, we would expect predictions based on P_2 to be both (1) computationally cheaper and (2) less accurate than predictions based on P_1. But note that, by hypothesis, R cannot be made on the basis of P_1. The calculation cannot be done. So the accuracy of prediction R on the basis of P_1 is at chance. Which means that, were P_2 an abstraction of P_1, one ought not be able to predict R, on the basis of P_2, at better than chance. If one can make such a prediction at better than chance then it follows that P_2 is not an abstraction of P_1.

Suppose that we predicate a normative property of some entity and, in virtue of so doing, can make predictions with better than chance accuracy. This gives us good reason to believe that the normative predicate tracks some external world properties. We want to know: is there any reason to believe that these properties are non-natural? Consider a sentence including an N-predicate and a sentence including a corresponding P-predicate. Suppose that the P-predicate is a complicated description of microphysical facts. If sentences including N-predicates are abstractions of sentences including the P-predicate, we have good reason to believe that sentences including N-predicates are made true by microphysical facts. Alternatively, if sentences including N-predicates are not abstractions of sentences including the P-predicate, we have good reason to believe that sentences including N-predicates *are not* made true by microphysical facts; i.e. the properties that are tracked by normative predicates are not token-token identical to microphysical properties. So long as comprehending application of the normative predicate in question allows us to make predictions of phenomenon at level of organisms, sentences including N-predicates *are not* abstractions of sentences about microphysical properties. Making predictions at the level of organisms in terms of microphysical facts would require more computational power than is available throughout the entire lifespan of the universe. The trade-off between computational and predictive accuracy that characterizes abstraction is broken. *If comprehending application of a predicate allows us to make better-than-chance predictions about phenomenon at, or above, the level of organisms, we have good reason to believe that the predicate tracks external world properties that are not token-token identical to microphysical properties.*

<u>But what about moral properties?</u>

Schmoral realism

We have now identified three distinct criteria a normative predicate must fulfill in order to help the non-naturalist's case.[25] (1) Comprehending application of the predicate must improve predictive power, (2) the best explanation of this fact is not available in terms of the predicate co-varying with mental states, and (3) comprehending application of the predicate must improve predictive power at the organismal level. If a predicate fulfills all of these conditions we have good reason to believe that comprehending application of the predicate co-varies with some set of non-natural properties. Suppose some moral predicate meets (1)-(3) above. Nothing I have said gives us any reason to think that the properties that co-vary with such a moral predicate are *moral* properties.

[25] Keep in mind the discussion of non-naturalism I offered at the outset of the chapter. As I'm using the term, the ethical non-naturalist is committed to the claim that moral facts are not token-token identical to microphysical facts.

Consider two examples I offered at the beginning of the chapter. Both “has phlogiston” and “is a witch” are predicates the comprehending application of which improved predictive power. But neither refers. There are no witches and there is no phlogiston. Nothing I have said so far gives us any reason to think that, even if moral predicates fulfill conditions (1)-(3) above, there are any moral properties.

As the relationship between ontology and semantics is generally understood, given some property, P, we know that P is a *moral* property just in case some of our moral predicates refer to P. The argument schema I have developed in this chapter says nothing about reference and, as such, by the lights of the traditional understanding of the relationship between ontology and semantics, the argument schema I have developed *cannot* demonstrate the existence of moral properties. At the end of the next chapter, after offering my considered defense of ethical non-naturalism, I will have more to say by way of a direct response to this objection. For now, instead of attempting to provide a response to the worry, I will argue that the objection need not concern us.

At the beginning of the second chapter of the dissertation I made a surprising claim: there are views that are correctly classified as anti-realist that I am willing to endorse. It may seem wildly incongruous to make this claim in a dissertation that putatively aims to offer a defense of ethical non-naturalism. This is a radical claim and requires some defense.

The fundamental question is: what aim does providing a successful defense of moral realism fulfill? An enormous portion of our lives revolves around moral considerations: What is the best kind of life I can live? What are my obligations to those around me? How should I act when I find myself in such-and-such a situation? Various versions of moral anti-realism threaten to undermine the centrality of these considerations to our lives. If the moral facts are nothing over and above preferences or social constructs then it seems that much of the importance we place on morality is misplaced. Moral realism offers to provide a defense of the importance we place on moral considerations; our moral concerns correspond to ontologically robust facts about our universe. A successful defense of moral realism promises to vindicate our fixation on living morally.

An analogy may help make clear what I have in mind. Suppose that Sam is a deeply religious man. Sam constantly strives to live his life in accordance with a set of what he thinks are divine commandments. If it turns out that there is no G-d, and consequently no divine commandments, there is an important sense in which Sam's choice of life project—the project of living by the strictures set out by divine decree—is unjustified.

One can imagine Sam coming across a constructivist about the divine. The constructivist tells Sam that, in fact, there are divine commandments. It just so happens that Sam has misunderstood the meaning of the word "divine." One does not need G-d for divine commandments. The divine commandments are social constructions. The constructivist tells Sam that his life project is not misplaced. Sam has successfully aimed at living in accordance with the strictures established by divine command.

It would be a surprise if Sam were to find any comfort in divine commandment constructivism. If Sam's life project is to be vindicated, there must be divine facts that are metaphysically independent of any human sentiments or practices. Analogously, I take it that only the existence of some set of mind-independent properties can vindicate our moral practices.

However, these mind-independent properties need not, technically, be *moral* properties. Meta-ethical accounts that are appropriately labeled "anti-realist" are sufficient to fulfill my broader aims. To see that this is the case, consider the following position. Start by supposing that our moral predicates fail to refer. That is, there are no moral properties. Nonetheless, suppose that comprehending application of our moral predicates co-varies with some set of ontologically robust properties. Further suppose that we have a theory of the relations that hold amongst these ontologically robust properties as well as a theory of the relations that hold between these ontologically robust properties and other kinds of properties, e.g. thoughts, actions, etc. Call the complete account of these intra-property and inter-property relations a *schmoral theory*. Further suppose that schmoral theory is isomorphic to some moral theory. That is, for any theoretical entity, T, that occupies some position, Q, in a moral theory, there is a corresponding entity, T', that occupies some position, Q', in a schmoral theory. Thus, though we use different words to describe the entities and relationships in a moral theory and a schmoral theory, for every entity and relationship in a moral theory, there is an isomorph in the schmoral theory. Technically, the view I have just sketched is a version of error theory. By hypothesis, there are no moral properties. Nonetheless, if we can vindicate a theory whose theoretical terms refer to mind-independent properties and the theory is isomorphic to moral theory, I take my project to have been successful.

Imagine that the obligatory nature of caring for our intimates would justify the importance I place on certain relationships in my life. Further imagine that there is a schmoral theory isomorphic to moral theory. In such a theory, there will be a theoretical term isomorphic to "obligatory." Call the property this theoretical term corresponds to the property of *being schmobligatory*. By hypothesis, if the property of *being obligatory* would justify the importance I place on certain relationships in my life, then the property of *being schmobligatory* would also justify the importance I place on certain relationships in my life.

Again, consider the analogy to our religious friend, Sam. For the sake of the analogy, suppose that a causal theory of reference is accurate.[26] "G-d" refers to G-d just in case (1) there was an original baptismal ceremony that successfully fixed the reference

[26] I take it that on nearly any reasonable story about reference, one will be able to construct a story similar to the one I am about to present. The strength of the analogy does not depend on the assumption of a causal theory of reference.

of the word "G-d" to the entity G-d and (2) Sam's use of the word "G-d" is appropriately causally related to this original baptismal ceremony. Further suppose that Sam finds out that the original baptismal ceremony failed to fix the referent of "G-d." The word does not refer. It follows that there are no divine commandments. It does not, however, follow that Sam's life project is fundamentally mistaken. We can imagine that, though the word "G-d" fails to refer, there is still an omniscient, omnipotent, omnibenevolent entity and this entity offered commandments that by-and-large share content with Sam's "divine commandments." So long as this is the case, Sam can be entirely unconcerned that, strictly speaking, there is neither a G-d nor divine commands. Sam has an accurate theory, isomorphic to his theory of divinity, which does all of the justificatory work he could wish for.

I take it that the following question is merely terminological: "are there moral properties?" I am largely unconcerned with terminological issues. I want to know if there are properties that fit a certain description. So long as the answer is "yes," it does not seem to matter much what we want to call these properties. If this means that the view I am defending is more accurately called "ethical non-naturalism*" than "ethical non-naturalism" or "schmoral realism" as opposed to "moral realism," so be it. So long as I have reason to believe that a certain kind of property, i.e. a set of properties matching a certain description, constitute an ontologically robust part of the universe, I take myself to have made good on the important question the moral realist asks.

A first objection to schmoral realism and a response

As I understand the moral realist's project, she can be content with a version of error theory so long as she has a theory of mind-independent properties that is isomorphic with a moral theory. Call this theory of mind-independent properties a *schmoral theory*. The following question becomes relevant: which moral theory does the schmoral theory have to be isomorphic with? If there were only one schmoral theory, or only one moral theory, theory isomorphism would be sufficient to vindicate the importance we place on moral considerations. But suppose that one could construct a theory of mind-independent properties isomorphic to *any* moral theory, no matter how implausible the moral theory is. If we were to see that we can "vindicate" *every* moral theory via isomorphism with a schmoral theory, it appears that we can't vindicate *any* moral theory via isomorphism. The "moral" theory that holds that torture is "obligatory" is incompatible with any plausible moral theory. A method that vindicates the one had better not vindicate the other. If both are vindicated then neither is; the "vindication" of both theories fails to offer any justification for my decision to have my life revolve around one theory, but not the other.

This worry need not be of any concern. While the view I am defending here is, technically, not a version of moral realism, the schmoral realist need be no more worried about a multiplicity of "moral theories" than need be the moral realist. The moral realist is committed to thinking that the theoretical terms of moral theory refer to mind-independent properties. One could then demand that the moral realist provide an answer to the following question: *which* moral theory? The moral realist need not be able to provide an answer to this question. Rather, she can pass the buck to the normative ethicist. The correct conceptual analysis of our moral predicates will provide us with our moral theory. All other "moral" theories are pretenders to the throne.

The schmoral realist can avail herself of the same strategy. Conceptual analysis of our moral predicates will reveal the connections that hold amongst the theoretical terms of our moral theory. A schmoral theory will be isomorphic to moral theory. Which moral theory? The moral theory that falls out of the analysis of our moral concepts.

One might still worry that schmoral realism is not sufficiently strong to capture the important commitments of the moral realist. I have talked about the analysis of *our* moral concepts. This way of talking seems to suggest that we all share a common set of moral concepts. One may doubt that this is true. It may seem more plausible to hold that predicates like "right" and "good," when used by me, express different concepts than the same predicates used by, e.g., a member of the KKK. If this is the case, one may wonder why the schmoral realist looks for a theory isomorphic to the theory that falls out of an analysis of the concepts *I* express when *I* use moral predicates, as opposed to a theory isomorphic to the theory that falls out of an analysis of the concepts the KKK member expresses when the KKK member uses moral predicates.

Again, I think that there is no reason for the schmoral realist to be worried. One might ask the moral realist the same question: what makes the properties referred to by your use of moral predicates, as opposed to the properties referred to by the KKK member's use of the moral predicates, the moral properties? The moral realist would likely respond with a puzzled look. The moral properties are the properties referred to by our use of moral predicates; this is just how the relationship between semantics and ontology works. But of course, there is no right way to use a certain set of phonemes. The KKK member can use the words "right" and "wrong" however she wants. Nothing of philosophical interest follows from the fact that people can use the same words to express different thoughts.

Just as the moral realist uses moral predicates to fix the moral properties, the schmoral realist uses moral predicates to fix the moral theory. The fact that people can use these same words to express different concepts poses no more threat to the schmoral realist's focus on moral theory than it does to the moral realist's focus on moral properties.

A second objection to schmoral realism

Schmoral realism is plagued by a further important and related problem. Just as a plurality of moral theories would cheapen the "vindication" of any given moral theory, a plurality of schmoral theories would cheapen the vindication of moral theory. Schmoral theory was supposed to vindicate moral theory by constructing an isomorphic theory out of *ontologically robust* properties. Success in constructing a schmoral theory would demonstrate that our moral concerns are not arbitrary, but are importantly related to the deep structure of the universe. However, suppose that there is a large number of schmoral theories corresponding to the moral theory. Such a multiplicity of schmoral theories raises the following question: which set of mind-independent properties are our moral concerns related to? Presuming that the schmoral theories have a sufficiently diverse set of theoretical entities, it does not seem that the schmoral theories have vindicated our moral theory.[27] It would appear that our moral concerns do not carve reality at its joints. Rather, our moral concerns are a kludge—responses to a mishmash of unrelated properties. The worry can be put another way. Schmoral realism promised to vindicate moral theory by showing that moral theory is isomorphic to a particular set of mind-independent properties. However, if there is a multiplicity of schmoral theories, we have failed to identify a particular set of mind-independent properties. Rather, we have identified a plurality of sets of mind-independent properties. It would appear that we lack any answer to the question, "which (mind-independent) properties justify the importance I have placed on living morally?"

Is there good reason to believe that there will be a plurality of schmoral theories? Unfortunately, the answer appears to be "yes." In the first chapter of the dissertation I discussed three different types of identities that might hold between some A and some B. A and B might be type-type identical, e.g. A may be the type *automobile* and B may be the type *car*. There are also token-token identities. If A and B are token-token identical then A and B are both tokens of the same property type. Token-token identity is strict,

[27] The existence of a multiplicity of schmoral theories is not immediately problematic. The problem only arises if the theoretical entities of the schmoral theories are sufficiently diverse. Presuming that the schmoral theories have by-and-large the same theoretical entities, we will have vindicated the core of moral theory. Minor disagreement along the edges need not be a cause for significant concern.

or numerical, identity. Lastly, there was disjunctive-type identity. In a disjunctive type identity, one rounded up a set of token identities, disjoined them, then set an identity between this disjunction and some property. In the first chapter, I presented the following example of this method:

> To get oneself a type-type identity one need only collect all of the instances of pain into an enormous (likely infinite) disjunction of particular pain instantiations and hold that this disjunction is identical to the property of *being in pain*. Let PI_n pick out an event that is token identical to an instance of pain. One can get a type identity for pain as follows: Pain = $\{PI_1 \vee PI_2 \vee PI_3 \vee \ldots \vee PI_n\}$.

Assuming that one is willing to allow disjunctive-type identities, there will likely be an infinite set of schmoral theories. Disjunctive-type identities allow for an unlimited amount of property gerrymandering. Once unlimited property gerrymandering is allowed, theory isomorphs are cheap.

One further condition

Vindication of a moral theory will require something stronger than mere isomorphism with a schmoral theory that relies on disjunctive-type identities. In addition to the three criteria I have already offered, we must add one last criterion that a predicate must meet in order to be of help to the ethical non-naturalist. This criterion will have to give us some reason to think that any predicate that meets it does not track a disjunctive-type property.

There are, broadly speaking, two types of disjunctive properties. A property might be finitely disjunctive or infinitely disjunctive. Consider a finite disjunctive property: *being compeiffel.* If some entity, E, is compeiffel, then E has the property of *being the Eiffel Tower* **or** *being a computer*. Consider the sorts of predictive power comprehending application of the predicate "is compeiffel" can get us. Suppose we accurately judge that *E is compeiffel.* We can now make all sorts of accurate predictions about E, e.g. E is not alive, E is an artifact, E is partially composed of metal, etc. Our predictive power even improves regarding those properties that the Eiffel Tower and a computer do not share, e.g. E executes computations, E is more than five stories tall, E was built for the world fair, etc. Suppose that we know that E is compeiffel. What is the highest probability with which we can predict that E is more than five stories tall? The answer is (the number of computers that are more than five stories tall + the number of Eiffel Towers)/(the number of computers + the number of Eiffel Towers).

Suppose we predicate P—a predicate that co-varies with a disjunctive property—of some entity E. We then attempt to make the prediction that *E is Q* based on our judgment that *E is P*. The probability of *E being Q* will be: (the number of entities ranged over by P that are Q)/(the number of entities ranged over by P).[28] As the number of disjuncts included in the disjunctive-type property goes to infinity, the denominator also goes to infinity. For most Qs, the numerator grows much more slowly, if at all.[29] Consequently, as the number of disjuncts included in the disjunctive-type property that co-varies with the comprehending application of P goes to infinity, we should expect the accuracy of our predictions based on comprehending application of P to drop to zero. We have a first blush response to worries about isomorphism with theories that rely on disjunctive-type properties. If comprehending application of a predicate improves predictive power, it may seem like we have reason to believe that the predicate does not co-vary with an infinitely disjunctive property.

[28] We of course, ought not take this equation to give us an accurate account of our ability to make predictions based on the comprehending application of the predicates in question. They do, however, appear to provide an upper bound on the accuracy of our predictions. After all, it would be shocking if the predictive power gained via comprehending application of a predicate outstripped the probabilistic relationships that hold between the predicted properties and the properties that co-vary with comprehending application of the predicate.

[29] The exception is if Q is the kind of property that everything has, e.g. the property of *having any properties*.

There is an obvious objection to this line of reasoning. Suppose that the comprehending application of the predicate “is a mammal” co-varies with the property of *being a mammal*. It immediately follows that comprehending application of the predicate “is a mammal” co-varies with the disjunctive property *being a mammal* ***or*** *being named Sam* ***or*** *being a reptile* ***or*** *being red*. So long as “is a mammal” co-varies with the property of *being a mammal*, it co-varies with any disjunctive property that has the property of *being a mammal* as one of the disjuncts. If the reasoning in the previous paragraph is accurate then “is a mammal” ought not improve our predictive power because the predicate co-varies with a disjunctive property with infinite disjuncts. But “is a mammal” does improve our predictive power. So something with the reasoning in the previous paragraph must be wrong.

Predictive power will only drop to zero in instances where comprehending application of a predicate co-varies with an infinitely disjunctive property and the predicate co-varies with *each* disjunct. By way of illustration, consider again our predicate “is compeiffel.” ‘Compeiffel’ co-varies with the disjunctive property *being the Eiffel Tower* ***or*** *being a computer*. Suppose, however, that the predicate “is compeiffel” does not co-vary with the property *being a computer*. If some entity, E, is compeiffel, then the probability of E being moral than five stories tall is equal to (the number of computers that are more than five stories tall + the number of Eiffel Towers)/(the number of computers + the number of Eiffel Towers). But if we know that comprehending application of the predicate “is compeiffel” co-varies with *being the Eiffel Tower* but **not** with the property of *being a computer*, it would be a miracle if comprehending application of “is compeiffel” improved our predictive power regarding computers. Thus, if we are trying to make predictions based on comprehending application of “is compeiffel,” we can safely ignore any disjunct that would improve our predictive power in virtue of tracking properties of computers. This allows us, when trying to get a rough idea of the kind of predictive power comprehending application of the predicate "is compeiffel" will get us, to simplify the probability calculation from (1) to (2):

1. (the number of computers that are more than five stories tall + the number of Eiffel Towers that are more than five stories tall)/(the number of computers + the number of Eiffel Towers)
2. (the number of Eiffel Towers that are more than five stories tall)/(the number of Eiffel Towers)

If we want to vindicate morality via theory isomorphism, we need some way to rule out isomorphism with a disjunctive property. Consider some normative predicate, P. Start by assuming that the predictive power gained via comprehending application of P is best explained by P co-varying with a disjunctive property. Further suppose that it is implausible to suppose that one can come up with a finite list of disjuncts such that co-variance with every member of this finite list of disjuncts could explain the range of predictive power we get from the comprehending application of P (setting aside the degenerate case of a disjunction whereby the disjunction is logically equivalent to a single disjunct, e.g. *being red* **or** *being red* **or** *being red*...). It follows that, if the predictive power of P is a consequence of P co-varying with a disjunctive property, P must co-vary with every disjunct of an infinite disjunction—the predictive power of P cannot be explained in terms of P co-varying with every disjunct of a finite disjunction. But as argued above, the predictive power of P cannot be explained in terms of the comprehending application of P co-varying with every disjunct of an infinite disjunction. Were this the case, we would not expect the comprehending application of P to improve predictive power. It follows that we must reject our original assumption. The best explanation of the predictive power gained by the comprehending application of P is not that P co-varies with some disjunctive property. With these considerations in mind, we can now introduce a fourth criterion a predicate must meet in order to be helpful for the non-naturalist.

Let me introduce a new technical term: "infinitely instantiable." A predicate is *infinitely instantiable* just in case (1) the comprehending application of a predicate improves predictive power and (2) co-variance with some finite set of disjuncts cannot explain the predictive gains garnered by comprehending application of the predicate. (2) will be met whenever there is no finite list of conditions sufficient for the comprehending application of a predicate.

Conclusion

My aim in this chapter was to develop an argument schema for vindicating the central commitments of the ethical non-naturalist. I have now done so. If a predicate fulfills the following three conditions, we have good reason to believe that it tracks a mind-independent non-natural property. (1) Comprehending application of the predicate must improve predictive power, (2) the best explanation of this fact is not terms of the predicate co-varying with mental states, and (3) comprehending application of the predicate must improve predictive power at the organismal level. Furthermore, if the predicate is infinitely instantiable, the properties that co-vary with the predicate constitute candidates for the construction of a theory isomorphic with a moral theory such that this isomorphic theory would constitute a vindication of the important commitments of the ethical non-naturalist. In the next chapter, I will attempt to apply this argument schema in defense of ethical non-naturalism. Before moving on to the next portion of the project, I would like to offer a concluding remark.

Insofar as there is an important insight in this dissertation, it can be found in this chapter. Behind everything that has happened in this chapter is the following train of thought. If one accepts the scientific worldview, one is committed to thinking that our best theories correspond with the structure of the universe. Theories are artifacts. Just like chairs and tables and books, theories are a part of our universe that can be approached from the scientific perspective. If our theories are isomorphic with some aspects of the structure of the universe we should expect the relations between the parts of our theories, i.e. pieces of language, to mirror the relations that hold between the entities our theories are about. Put simply, if our language mirrors the universe then we should expect to be able to learn things about the universe by the study of our language. Furthermore, our language *is* part of the world that we can study via the methods of science. Thus, there is some reason to hope that the application of the methods of science to our language can tell us something about the deep metaphysical structure of the world we inhabit. The arguments I have offered in this chapter constitute one way one might go about trying to make good on this train of thought. Even if everything else I have said in this chapter is unconvincing, I hope that this rough sketch of the motivations behind my argumentative strategy is independently interesting. One of the central methodological questions in contemporary philosophy is: to what extent can empirical evidence inform the philosophical project? Something like the above considerations might offer to provide a middle ground between the philosopher's traditional methods, heavily reliant on language, and the empirical methods of the youngest generation of philosophers.

CHAPTER FIVE:
A METHODOLOGICALLY NATURALIST DEFENSE OF ETHICAL NON-NATURALISM

Introduction

In the previous chapter I argued that, if a predicate fulfills the following set of conditions, it may be able to serve as a starting point for a non-naturalist friendly development of a theory isomorphic to a moral theory: (1) Comprehending application of the predicate improves predictive power; (2) these gains in predictive power cannot be explained by comprehending application of the predicate co-varying with anyone's mental states; (3) the predicate improves predictive power at the organismal level; (4) the predicate is infinitely instantiable.

This chapter is dedicated to making the case that there is some predicate that fulfills these four conditions. Further, my aim is to demonstrate that one can construct a theory isomorphic with moral theory based on the properties that co-vary with such a predicate. If I can be successful in these two tasks, I may not have offered a methodologically naturalist defense of ethical non-naturalism; however, I will have offered a methodologically naturalist defense of a meta-ethical account that captures the important features of ethical non-naturalism.

What predicates are promising?

Before jumping headfirst into the project of attempting to identify some range of predicates that can ground a theory isomorphic with a moral theory, it would be wise to have some strategy for identifying those predicates that are likely to be of help. Which predicates are likely to be helpful? Our aim is to build a theory isomorphic with a moral theory. Thus, if a moral theory posits *being good* as a theoretical property, our theory must posit some isomorphic property. This immediately suggests a promising starting place to look for predicates that may be helpful: normative ethical theories. If we can show that the comprehending application of the primary moral predicates in a normative ethical theory co-vary with mind-independent non-natural properties, we can construct a theory isomorphic with a moral theory simply by relying on the relations that hold between these co-varying properties. My strategy in this chapter will be to consider a range of predicates associated with moral theories.

At the end of the previous chapter I noted that the strategy of finding a theory isomorph is technically too weak to count as a defense of ethical non-naturalism. I will be returning to this issue at the end of this chapter; however, for the time being I want to make a terminological fiat. In what remains of this chapter I will talk as if finding a theory reliant on mind-independent non-natural properties isomorphic to moral theory is sufficient to have provided a successful defense of ethical non-naturalism. While this kind of talk is inaccurate, the pretense will allow me to avoid having to repeatedly put my points in precise, but difficult to read, technical language. So long as this caveat is kept in mind, the gains in readability justify the fact that I will make a series of claims that are, technically, false.

Organismal level predicates

In the previous chapter I laid out four conditions a predicate must meet if it is going to help establish ethical non-naturalism: (1) Comprehending application of the predicate improves predictive power. (2) These gains in predictive power are not best explained by comprehending application of the predicate co-varying with anyone's mental states. (3) The predicate improves predictive power at (or above) the organismal level. (4) The predicate is infinitely instantiable. As I understand the distinction between ethical naturalism and ethical non-naturalism, the non-naturalist is committed to the rejection of token-token identities between moral facts and physical facts. The third condition—that comprehending application of the predicate improves predictive power at (or above) the organismal level—does all of the work in establishing that the facts that co-vary with the comprehending application of some predicate are not token-token identical with microphysical facts. In the remainder of the chapter, I will essentially ignore this third condition. For each predicate I consider, I will have to show that comprehending application of the predicate improves predictive power. In so doing, I will offer examples that demonstrate that, if comprehending application of the predicate in question improves predictive power, it does so at (or above) the organismal level. In what follows, I will not take the time to note that comprehending application of predicates improves predictive power at (or above) the organismal level. Instead, I will rely on the reader to note that this is the case.

First attempt: "good" and "bad," "right" and "wrong"

The natural place for the non-naturalist to start is with our preferred moral predicates: "is good" and "is bad," "is right" and "is wrong." There is an immediate advantage to starting with these predicates. If we can use the argument schema I developed in the fourth chapter to show that any one of these predicates co-varies with a mind-independent non-natural property, we will immediately be in a position to construct a theory isomorphic to moral theory; "good," "bad," "right," and "wrong" are the fundamental building blocks of normative theory. There are, however, reasons to be doubtful that this approach will be successful. The trouble is that, though one could surely construct moral theory out of the properties that putatively correspond with these predicates, it is not clear that these predicates can successfully instantiate the argument schema I offered in the fourth chapter; it is plausible to think that any predictive power one gets by applying these predicates can be explained via co-variance with mental properties.

Suppose that some action, A, is wrong or that some state of affairs, S, is bad. What kind of predictions does knowing that *A is wrong* or that *S is bad* allow one to make? One might think that having this kind of knowledge will increase one's ability to predict how people will act. Knowing that *A is wrong* or that *S is bad* might help one predict that agents will be motivated not to do A or to try and prevent bringing about S. The problem is that making the judgment that *A is wrong* or *S is bad* doesn't, by itself, tell us anything about how others will behave. In order to predict behavior, one must know that some agent believes that *A is wrong* or believes that *S is bad*. It is difficult to see how the wrongness of A could possibly influence an agent's behavior in the absence of the belief that *A is wrong*. Furthermore, if we know that an agent believes that *A is wrong*, it is very difficult to see how accurate judgments about facts regarding the wrongness of A could improve our predictive power. Suppose we know that Sam believes that *A is wrong* and we further know that *it is not the case that A is wrong*. Knowing the latter does not appear to change our behavioral predictions at all. Once we know that Sam believes that *A is wrong*, the rightness or wrongness of A appears entirely superfluous to our ability to predict how Sam will act. It seems that knowing that *A is wrong* fails to offer any improvements in our ability to predict behavior over and above the predictive improvements we get by comprehendingly applying folk psychological predicates. This is exactly the point that Harman pushes in the argument we considered in chapter three (Harman). The best explanation of any predictive gains regarding behavior we get by comprehendingly applying "right/wrong" and "good/bad" appears to only require that these predicates co-vary the mental states of certain agents. Consequently, I will have to look somewhere else to find the kind of predicates the non-naturalist needs.

<u>Second attempt: character traits</u>

The failure of the first attempt to locate appropriate predicates does not suggest that the ethical non-naturalist ought to give up. If the argument schema I developed in the fourth chapter cannot be used, straightforwardly, to show that "is good" and "is bad" or "is right" and "is wrong" co-vary with the right kind of properties, the non-naturalist can attempt to find a more roundabout way to vindicate her view. As I understand the commitments of the ethical non-naturalist, she must hold that there is some set of mind-independent non-natural properties from which one can construct a theory isomorphic to moral theory. This does not, however, require that we use "is good" or "is right" as the fundamental building blocks of our theory. A toy example will help to illustrate. It is useful to think about normative ethical views in terms of the moral property that they take to be primary. The consequentalist takes the good to be prior to the right, i.e. the consequentalist thinks that rightness can be analyzed in terms of goodness. The deontologist tends to think that the right and the good are on par; neither can be analyzed in terms of the other (Ross). On at least some versions of virtue ethics, virtue is the primary moral property (see, e.g. [Hursthouse]). One can imagine a view whereby various virtues are taken to be non-natural. Rightness can then be defined in terms of virtue: given some situation, S, the right action is any action a fully virtuous person would take in S. One could also offer an analysis of good in terms of the virtues. One is beneficent just in case one has a standing disposition to act in ways that maximize the good. Generally, we think that goodness is primary, thus we think that the above claim offers an analysis of beneficence in terms of goodness. But nothing forces us to understand the claim this way. We can reverse the direction of analysis and offer an analysis of the good in terms of beneficence: the good is that which the beneficent has a standing disposition to maximize. I do not take this sketch of a view to be particularly defensible; however, it nicely illustrates the ethical non-naturalist's strategy given the difficulties involved with applying the argument schema I developed in chapter four to the predicates "is good" and "is right." If one can, relying on the properties that co-vary with virtue predicates, build a theory with properties isomorphic to *being right* and *being good*, then one can be an ethical non-naturalist while thinking that neither the predicates "is right" or "is good" fulfill the four conditions I have established.

One can find, in various kinds of virtue ethical approaches, a variety of candidate predicates for the non-naturalist's project. I will start by considering attempts to vindicate moral realism that stem from attempted responses to Harman's critique of moral realism. Work by both Russ Shafer-Landau ("Moral and theological realism") and Nicholas Sturgeon suggests that virtue predicates may offer a promising route for the non-naturalist. The virtue ethicist attempts to analyze rightness/wrongness in terms "thick" moral concepts, e.g. courageous, honest, kind, etc. (Williams). In other words, the virtue ethicist attempts to analyze rightness/wrongness in terms of the *character traits* of a virtuous individual (Hursthouse).

For this strategy to be effective, it must be the case that (1) the comprehending application of character trait predicates improves predictive power, (2) this predictive power cannot be explained in terms of co-variance with mental properties, and (3) character trait predicates are infinitely instantiable. Consider the following, by way of a starting place for making good on (1)-(3), on offer from Shafer-Landau:

> Many nonmoral facts are counterfactually dependent on moral facts. If certain moral facts had not obtained, other nonmoral facts would not have obtained… Suppose that we invoke a moral fact to explain a nonmoral facts. The employee pension fund is now drained; what accounts for this? It can be perfectly natural to cite the venality, greed and moral corruption of the corporate executives who perpetrated the fraud… If we kept everything else fixed, but assumed that the executives were not greedy, venal and corrupt, then the outcome would have been quite different—the pension fund would presently be well-endowed… But the pension fund would remain depleted in the nearest possible world, one in which the specific natural facts that constituted the corruption were changed only very slightly… ("Moral and theological realism" 317-320)

Here, Shafer-Landau attempts to give a methodologically naturalist defense of ethical non-naturalism. His strategy is not the same as mine; whereas I want to argue that mind-independent non-natural moral properties play a role in the best explanation of the predictive power provided by moral predicates, Shafer-Landau argues that mind-independent non-natural moral properties play a role in the best explanation of behavior, i.e. the best explanation of the executive's theft is that she was greedy. Despite the differences in our strategy, Shafer-Landau's argument appears to nicely illustrate two facts about character traits.

First, knowing facts about character traits appears to provide predictive power. Imagine one receives a questionnaire with questions like the following: "Sarah is an executive at such-and-such a company. Sarah is venal, greedy, and morally corrupt. Has Sarah stolen money from the company's pension fund?" One can either answer "yes" or "no." It seems *very* plausible to suppose that knowing various character traits of the executive in question will allow one to successfully answer the questions at better than chance.[30] If we know nothing about Sarah we merely have to guess whether or not she has stolen money from the company. Alternatively, if we know that Sarah is greedy or, alternatively, that Sarah is honest, it seems likely that we will be able to do better than merely guessing. After all, honest people aren't exactly the type to steal money.

Second, it looks like character traits are infinitely instantiable. Shafer-Landau is surely right that "the pension fund would remain depleted in the nearest possible world, one in which the specific natural facts that constituted the corruption were changed only very slightly" ("Moral and theological realism"). Perhaps Sarah stole the money merely by straightforwardly withdrawing it from the bank. Alternatively, maybe her method of theft involved various complicated schemes for covering her tracks. Of course, Sarah need not have stolen money from the pension fund in order to be greedy. Perhaps Sarah has not stolen any money from the pension fund; however, she has cut workers' pay so

[30] Remember that the comprehending application of a predicate need only offer minor improvements in predictive power to give us reason to believe that a predicate co-varies with a property. We may need to know a lot more about Sarah to have much chance of accurately predicting if Sarah stole money from the pension fund. For example, Sarah may have a morbid fear of prison that prevents her from doing anything illegal. We are not, however, after accurate prediction. We are merely looking for better than chance predictive accuracy.

that she can increase her already hefty compensation. There is an indefinite number of ways that Sarah could instantiate her greediness.

The same can be said of the other virtues. Consider, e.g. courage. There is an infinite number of scenarios in which one can display courage. In each instance, one's courage is instantiated differently. One can be courageous in the face of physical danger, one can be courageous in the face of social pressure, one can be courageous in scenarios where one has very little at stake and one can be courageous in scenarios where one has everything at stake. For any finite list of conditions sufficient for instantiating courage, there will always be some condition sufficient for the instantiation of courage not on the list.

Finally, it needs to be shown that the predictive power gained by comprehending application of character trait predicates cannot best be explained via co-variance with mental properties. I remain unconvinced regarding the prospects for successfully showing that this is the case. Much depends on whether one thinks that character traits are something over-and-above mental states. One might think that *being greedy* is nothing more than having a certain set of mental states. Alternatively, one might think that *being greedy* consists of a range of behavioral dispositions that come apart from any description of an individual's mental states. Happily, I can leave this question unanswered. Despite what I have said so far, there are good reasons to be doubtful that comprehending application of character trait predicates can improve predictive power.

Empirical evidence suggests that the application of character trait terms does not, in fact, increase one's predictive power. Doris has relied heavily on work from the situationist research program in psychology in an attempt to argue that virtue ethics is not tenable. Given that the second attempted defense of ethical non-naturalism is tightly tied to virtue ethics, it should come as little surprise that Doris' arguments are relevant here. I do, however, intend to divorce myself from Doris' work *qua* critique of virtue ethics. Independent of the accuracy of Doris' intended worry, his work is relevant to this attempt to vindicate ethical non-naturalism via character traits.

I quote, at length, Doris's summary of the findings of the situationist research program:

> [C]haracter is expected to have regular behavioral manifestations; we believe that the person of good character will behave appropriately, even in situations with substantial pressures to moral failure, and we are similarly confident that we would be foolish to rely on the person of bad character. This interpretive strategy presupposes that the attribution of a character trait allows us to predict an individual's behavior in novel circumstances; we may not have previously observed Jim's behavior on a foundering ship, but if we know he is courageous, we know that he will perform his office properly should such a situation arise. Unfortunately, experimental evidence… suggests that this approach, however commonplace it may be, is inadequate to the facts of actual behavior: trait attribution is often surprisingly inefficacious in predicting behavior in particular novel situations, because differing behavior outcomes often seem a function of situational variation more than individual disposition. To put this crudely, people typically lack character…
>
> Whatever behavior reliability we do observe may be readily short-circuited by situational variation: in a run of trait-relevant situations with diverse features, an individual to whom we have attributed a given trait will often behave inconsistently with regard to the behavior expected on attribution of that trait. (Doris 505-507)

These findings are well confirmed (consider, e.g., some seminal articles: [Haney, Banks, and Zimbardo; Isen and Levin; Milgram]). Even if the situationists are wrong, in light of the evidence they have compiled, we ought to be wary of offering a virtue ethical defense of ethical non-naturalism. That situationism is a robust and respected research tradition in psychology is enough reason, at least for the time being, to be doubtful that character trait attribution improves predictive power. This is, in turn, enough reason to avoid deploying this virtue ethical strategy in defense of ethical non-naturalism.

Third attempt: functional properties

Contemporary virtue ethics is a far cry from its Aristotelian predecessor. Though inspired by Aristotle, it would be a stretch to say that contemporary virtue ethics is

Aristotelian. The fundamental difference lies in the centrality of a *telos*, or function. Aristotle thought that (biological) species could be classified by their function. Every member of a given species shared a *telos* with her conspecifics; humans are no exception (Aristotle). On the Aristotelian view, human virtues were those character traits that were conducive to, or perhaps constitutive of, the achievement of the human *telos* (Aristotle). The notion of a biological *telos* (at least of the Aristotelian variety) has become scientifically suspect. For this reason, contemporary virtue ethicists have moved away from Aristotle's original account of virtue. In a somewhat surprising move in a methodologically naturalist account, the non-naturalist might attempt to re-introduce the notion of a function in defense of non-naturalism.

Instead of looking to biology for a function—commitment to methodological naturalism seems to rule this out—the non-naturalist can look to social roles. Consider the following roles: teacher, bus driver, chef, librarian, parent, etc. Occupation of each role comes along with certain demands on excellence, e.g. one can easily be a *good* father without understanding the Library of Congress filing system; however, the same cannot be said of a librarian. When fulfilling each role well, one is thereby performing some function. The chef performs the function of cooking, the teacher the function of teaching, and the bus driver the function of driving a bus. In each case, one is a good specimen of one's functional type if and only if one fulfills one's function well. That is, a good teacher is a teacher who teaches well. A good chef is a chef who cooks well. A good bus driver is one who drives well. Call these role specific *teloi* "role properties" and call the corresponding predicates "role predicates."

In order to fulfill the anti-reductionist argument schema I offered in the previous chapter, "is good" as predicated of a functional role must (1) improve predictive power, (2) this improvement in predictive power cannot be best explained by co-variance with mental properties, and (3) the predicates must be infinitely instantiable.

It seems clear that the comprehending application of role predicates will improve one's predictive power. If one knows that Sam is a good chef, one is in a position to accurately predict that the food Sam cooks will taste good. If one knows that Ian is a good bus driver, one is in a position to predict accurately that Ian does not often run into other vehicles with his bus. If Richard is a good father one is in a position to predict that Richard's children are unlikely to feel starved for love. Role predicates are the first normative predicates we have considered that appear to genuinely improve one's predictive power.

It appears to be improbable that the improvement in predictive power offered by the comprehending application of each role predicate is best explained in terms of co-variance with mental properties. Consider the predicate "is a good teacher." What mental property, or set of mental properties, is it plausible to think could explain the predictive power gained by comprehending application of the predicate? Consider some of the predictions comprehending application of the predicate appears to let us make: students will have a strong grasp on the material, students will show significant improvement in critical thinking skills, students will become more intellectually curious, etc. It is deeply implausible to suppose that the mental states of a teacher can offer a complete explanation of these kind of changes in his students. The mental states of a teacher cannot directly influence the mental states and habits of students. It follows that the mental states of a teacher cannot adequately explain the predictive gains achieved via comprehending application of the predicate "is a good teacher." The same style of reasoning will work for a broad range of role predicates.[31]

[31] It strikes me as probable that role properties can be analyzed in terms of behavioral dispositions. À la the arguments in the fourth chapter, behavioral dispositions are not token-token identical with microphysical facts. Consequently, role properties can be non-natural properties even though they are analyzable in terms of behavioral dispositions.

Are role predicates infinitely instantiable? The answer seems to be “yes.” Consider what it is to be a good teacher. One might think that a good teacher is any teacher who gets his students to learn material. There is an infinite number of ways that one might effectively get students to learn material. Every teacher does things a little differently. Every teacher who effectively gets students to learn material instantiates a different set of properties that make it appropriate to apply the predicate “is a good teacher.” Effectively getting students to learn material is not the only way to be a good teacher. A teacher might be abysmal at teaching material; however, he may also be fantastic at inspiring intellectual curiosity in his students. Or perhaps while poor at inspiring intellectual curiosity in his students, he is great at making students consider the real world relevance of the material they are being taught. Given any finite list of sufficient conditions for the comprehending application of the predicate “is a good teacher,” it seems that one will be always be able to find some cluster of properties, not already on the list, which would allow for the comprehending application of the predicate. I take it that the same can be said of wide range of role predicates. It is hard to imagine a finite disjunctive property such that comprehending application of role predicates co-varying with each member of the disjunct can offer an adequate explanation of the predictive gains offered by comprehending application of role predicates.

This second Aristotelian strategy has significant promise. It appears that role predicates fulfill all four of the conditions set out in the fourth chapter of the dissertation. This is, by itself, a striking result. One might have thought that commitment to methodological naturalism would eventually push one in the direction of renouncing the existence of any mind-independent non-natural normative properties. This turns out not to be the case. It appears that role predicates co-vary with ontologically robust properties. This consequence of methodological naturalism is significant and worthy of attention unto itself. Unfortunately, my aim is not to demonstrate the existence of just any mind-independent non-natural normative properties. It is my goal to show, specifically, that there are non-natural *moral* properties—or at least properties that are isomorphic with putative moral properties. It is not enough to show role properties fulfill all four conditions I set out in the previous chapter. I must further show that a moral theory can be built from the mind-independent non-natural normative properties I have identified. It is not clear that role properties, by themselves, can be used to arrive at a moral theory.

In presenting this third attempted defense of non-naturalism, I carefully cherry picked the functional roles I used as examples: teacher, parent, bus driver, etc. It seems like it's a good thing to be a good teacher, or a good parent, or a good bus driver.[32] Not all functional roles work this way. Consider the following: slave driver, dictator, assassin, and concentration camp commandant. In the same sense that one can be a good teacher, one can be a good assassin. While it is a good thing to be a good teacher, it is a bad thing to be a good concentration camp commandant. One ought not strive to be a good concentration camp commandant. Much the opposite. If one is to arrive at an isomorph to moral theory based on role predicates, one must be able to give some account of why being a good teacher is something worth aiming at whereas being a good slave driver is not something worth aiming at. Put another way, in order to construct an isomorph to moral theory based on role predicates, one must have some way of sorting the good functional roles from the bad functional roles.

Aristotle had a way to do this: the valuable functional roles were the ones that helped fulfill the human *telos*. To mirror Aristotle's strategy, we need to identify a meta-

[32] On further reflection it is implausible to think that it is always a good thing to be a good teacher or a good bus driver. Imagine a good teacher who is tasked with teaching his students racist propaganda. All things considered, it would be better were this teacher bad at teaching.

telos, i.e. a *telos* such that the instantiation of the right kind of role properties—teacher, father, etc.—helps one achieve the *telos* and the instantiation of the wrong kind of role properties—slave driver, thief, etc.—hinders one from achieving the *telos*. Aristotle relied on the human *telos*, where this *telos* was understood biologically. Biological *teloi* are suspect. The ethical non-naturalist who is inclined in the direction of methodological naturalism cannot follow Aristotle in using the human *telos* as a sorting mechanism for the other *teloi*.

One might be tempted to rely on a normative, and not a biological, *telos* to play the role of *teloi* sorter. Thus, one could replace the property of *good human* with the role property of *good person*. For this strategy to work, the property of *being a good person* must be mind-independent and non-natural; the property of *being a good person* would be serving as the cornerstone for the ethical non-naturalist's construction of an isomorph to moral theory. In order to demonstrate that "good person" tracks a mind-independent non-natural property, one must be able to show that the comprehending application of "good person" provides improved predictive power. What might the comprehending application of "good person" help us predict? The most promising route is to claim that it can help us predict behavior. If I know that Ben is a good person, perhaps I can predict at better than chance accuracy that, e.g., he will help an old lady get her groceries across the street. But this strategy for defending the predictive power of the comprehending application of a predicate looks familiar. We previously considered a similar move with more fine-grained character trait attributions, e.g. "honest," "greedy," etc. The situationist challenge applies to "good person" as much as it does to "honest person" or "greedy person." If the situationist is correct, we should not expect high level character trait attributions like "good person" to offer any improvements in predictive power.

There is a further, conceptual challenge, to the strategy of using "good person" as a sortal for the other *teloi*. The trouble is that there appears to be a fundamental conceptual link between *being a good person* and the other *teloi*. *Being a good person* offers to act as a sortal because successfully fulfilling the right kind of role properties, e.g. *being a good father*, *being a good teacher*, etc., is constitutive of *being a good person* whereas *being a good assassin* is not partially constitutive of *being a good person*. Thus, introduction of the meta-*telos being a good person* allows us to explain why one should aim at *being a good father* but one should not aim at *being a good assassin*. The very aspect of this putative meta-*telos* that allows it to work as a sortal suggests it cannot play the role the ethical non-naturalist needs it to play. If possessing the right kind of *teloi* is constitutive of *being a good person* then it seems that "being a good person" can be analyzed in terms of the other *teloi*. If "being a good person" can be analyzed in terms of the other *teloi*, it can't function as a sortal. To know what counts as *being a good person* we already need to know which *teloi* are partially constitutive of *being a good person*. Which is to say, we already need to be in possession of a method for sorting the right kind of *teloi*—e.g., *being a good teacher*, *being a good father*—from the wrong kind of *teloi*—e.g., *being a good assassin*, *being a good slave driver*.

Fourth attempt: "good for"

We have, thus far, rejected two neo-Aristotelian defenses of ethical non-naturalism. Perhaps surprisingly, we have not yet exhausted all such neo-Aristotelian strategies. In a provocative article, Evan Fales offers another way one might attempt to deploy *teloi* to ground an isomorph to moral theory. He argues that living organisms are *intrinsically teleologically organized systems* (henceforth, ITOS). In virtue of being teleologically organized, there are things that are *good for* and *bad for* living organisms. It is Fales' goal to develop a meta-ethics that "grounds truths about moral permission and obligation on truths about what is good or bad, and [that] takes goodness and badness always to be goodness or badness *for* a certain being or type of being" (Fales 16). This suggests a third strategy the non-naturalist could pursue. Perhaps the argument schema I developed in the previous chapter can be successfully instantiated by some instances of the predicates "is good for" and "is bad for." The non-naturalist could then attempt to build an isomorph to moral theory on the basis of these properties. Again, there are three key questions: (1) does comprehending application of the predicates "good for" and "bad for" improve predictive power, (2) is this improvement in predictive power best explained by co-variance with mental properties, and (3) are the predicates "good for" and "bad for" infinitely instantiable?

It seems to be the case that the comprehending application of the predicate "good for" and "bad for" improves predictive power. Start with a very simple case. Suppose that we know that *such-and-such is bad for plants*. We also know that *such-and-such* has been introduced into the environment of some plant. In virtue of having these two pieces of information, it looks like we are in a position to make the following predictions with better than chance accuracy: the plant will grow more slowly than before the introduction of *such-and-such* into the environment *and* the leaves of the plant are more likely to turn brown than before the introduction of *such-and-such* into the environment. Similar predictions can be made about people. Suppose that I know that ingesting *such-and-such* is bad for Sam and I know that Sam has recently ingested *such-and-such*. This seems to put me in a position to predict, at better than chance, that Sam will, at some point, regret having eaten *such-and-such* or that Sam will soon feel ill or vomit. Some comprehending applications of the predicates "good for" and "bad for" appear to improve predictive power.

Furthermore, it is deeply implausible to suppose that the predictive power gained via the comprehending application of the predicate "is bad for" is best explained in terms of co-variance with mental properties. Comprehending application of the predicate allows us to predict such things as Sam vomiting or a plant growing more slowly. These are not the kind of phenomena that, generally speaking, are caused by mental states. It follows that it is implausible to suppose that the predictive power gained via comprehending application of the predicates “is good for” and “is bad for” is, at least in these sorts of cases, best explained by co-variance with mental properties.

So far, so good. Are the predicates “good for” and “bad for” infinitely instantiable? The answer appears to be: “yes.” Consider an example. Were Sam to drink bleach, right now, it would be bad for him. Is drinking bleach always bad for Sam? No. If Sam has recently eaten rat poison, drinking bleach may well save his life; drinking bleach makes one vomit and vomiting rat poison is usually a good thing. Of course, one could further suppose that Sam ate the rat poison in order to save his life from some other condition, e.g. an imminent heart attack caused by blood that clots too easily. If this is the case, drinking the bleach would, once again, be bad for Sam. The upshot: there does not seem to be any finite list of sets of properties that can capture all of the sufficient conditions for the comprehending application of the predicates “is good for” or “is bad for.” Given any finite list of realizers the presence of which would be sufficient for the comprehending application of the predicate “is good for Sam” and “is bad for Sam,” one will always be able to cook up a set of properties, not already on the list, that would allow for comprehending application of the predicate.

This gives us good reason to believe that the predicates “is good for” and “is bad for” can successfully instantiate the argument schema I developed in the previous chapter. This is the second successful application of the anti-reductionist argument schema I offered in chapter four. Can an isomorph to moral theory be constructed out of the properties that co-vary with the predicates “is good for” and “is bad for?” I am skeptical.[33]

[33] It is worth saying a bit more about what requirements I think a plausible normative ethical theory must fulfill. It is my hope that most of what I have to say here is largely uncontroversial. I take it that, in order for a normative ethical theory to be plausible, it must capture a sizeable subset of our pre-theoretic considered moral judgments. I have in mind here something like Rawl’s narrow reflective equilibrium (Rawls 2001). I tend to understand normative ethical theorizing on the general model of theory construction. A good theory is one

If we are going to construct an isomorph to moral theory out of the properties that co-vary with the predicates "is good for" and "is bad for," we must be able to save a significant subset of our considered moral judgments by showing that they can be derived from a meta-ethics that takes these properties to be fundamental. There is reason to doubt that we will be able to do so. The trouble is that the predicates "is good for" and "is bad for" are relative. Something can be *bad for Sam* but *good for me.* Consider Sam's pain. I take it that Sam's pain is *bad for Sam*. I also take it that, on any plausible normative ethical view, it is impermissible to cause Sam excruciating pain without some good reason for doing so. If one thinks that pain is bad *simpliciter*, it is clear why, *ceteris paribus*, I ought not cause Sam excruciating pain: by doing so I am making the universe a worse place. But suppose that one cannot help oneself to the notion of good and bad *simpliciter*. One cannot argue that I ought not cause Sam excruciating pain because doing so makes the universe a worse place. This would be to suppose that some non-relative predicate involving "good" and "bad" fulfills the four conditions laid out in chapter four. Given the predicates I am relying on, one can only hold that causing Sam excruciating pain makes Sam's life less valuable. But unless one can help oneself to a non-relativized notion of "goodness" or "badness," there is no way to move from the claim that *Sam's excruciating pain makes Sam's life less valuable* to the claim that *ceteris paribus, it is morally impermissible for me to cause Sam excruciating pain.*

that, minimally, can accommodate most of the evidence. When doing normative ethics, our considered judgments constitute our evidence. If a normative ethical theory fails to capture a sizeable subset of our pre-theoretic considered moral judgments we have good reason to believe either that (1) the theory has changed the subject or (2) the theory is not very good. This is all very broad strokes. I mean for it to be. By leaving nearly all of the details of the view unspecified, I hope to have painted an inclusive enough picture of normative ethical theorizing that there is little room for controversy.

Fales has suggested a potential route for defending the strategy (personal communication, date unknown). The key move in this proposed defense is accepting that humans are fundamentally social beings. In virtue of being interconnected, *ceteris paribus*, if something is bad for Sam then it is also bad for me. If everyone around me is ill, or unhappy, or malnourished, *ceteris paribus*, I will be less well off than if those around me were healthy, happy, and nourished. If I frequently engage in normal human interactions with Sam, if Sam is ill, unhappy, or malnourished, I can expect these interactions to be greatly degraded. Thus, *ceteris paribus*, if something is bad for Sam, it is bad for me.

This line of reasoning, if accurate, tells us something about the comprehending application conditions of the predicates "is good for" and "is bad for" as well as the comprehending application conditions of the predicates "is good" and "is bad." In particular, it tells us that the comprehending application conditions of the two types of predicates are intimately related. It follows that the properties that determine when each predicate is comprehendingly applied are interrelated—a fact that gives us some reason to think that, were we to create a theory of the properties that co-vary with the comprehending application of each of the predicates, the isomorph of the properties *being good* and *being bad* would be related to the isomorph of the properties *being good for* and *being bad for*. Examination of the relationship between the comprehending application of these two kinds of predicates can tell us something about what relations would hold between the theoretical entities in the theory we hope is isomorphic to moral theory.

In conversation, Fales expressed some hesitance about this suggested sketch of a defense of a normative ethics constructed from *being good for* and *being bad for*. I am also somewhat doubtful that reliance on the social nature of humans can ground a successful defense of this most recent attempt to construct an isomorph to moral theory. Consider, again, the claim with which I ended the previous paragraph: *ceteris paribus*, if something is bad for Sam, it is bad for me. The problem is that the *ceteris paribus* conditions are too easily broken. If Sam and I interact daily under normal circumstances, then I am inclined to think that, if something is bad for Sam, it is likely also (at least somewhat) bad for me. But it is easy to concoct a scenario where the *ceteris paribus* conditions do not hold. Suppose I am placed in front of two buttons. Pushing the button on the right will cause excruciating pain to someone I will never meet. Pushing the button on the left will cause a prolonged state of ecstasy in this same person. If the reader would like, we can further stipulate that no one will ever know (1) which button I pushed or (2) even that I was put in the situation. We can even imagine that I will fail to remember the entire affair. It seems clear that the right thing to do is to push the button that will cause prolonged ecstasy. But it is very difficult to see how my choice to push one button or the other would be either *good for me* or *bad for me*. I will never meet the person whose life I am influencing. Relying only on the properties of *being good for* and *being bad for*, it is difficult to see how to ground the claim that I ought to push the button on the left. Given that my choice to push a particular button appears to be neither good for me nor bad for me, it seems that any normative ethics grounded in these two properties is committed to holding that I ought to be indifferent regarding which button I push; however, this is surely the wrong result. Consequently, it is doubtful that any theory based on the properties that co-vary with the predicates “is good for” and “is bad for” will be isomorphic to moral theory.

One does not need to look to such exotic thought experiments to raise related worries. The *ceteris paribus* conditions appear to be regularly broken in all sorts of familiar social conditions. Imagine that I live in a racist or homophobic society. In such a society, various overt acts of racism or homophobia can garner social favor whereas doing the right thing, i.e. standing up for oppressed minorities, could lead to social sanction or, as is far too often the case, physical harm. Even though it would be bad for me to stand up for oppressed minorities, and good for me to embrace the role of oppressor, surely the former is the right thing to do and the latter is not.

One does not even have to go as far afield as considering racist or homophobic societies to make the point. The tendency towards ingroup/outgroup behavior has been well documented (see, e.g, [Tajfel]). It's plausible that picking on the unpopular kid in elementary school can go a long way towards making any given child part of the in-group. Given the often brutal nature of pre-adolescent social interactions, it may well be good for any individual child to bully the unpopular kid, though it's surely bad for the unpopular kid to be bullied.

Before concluding this section, one final move needs to be ruled out. It may be tempting to argue that, in an important sense, being a schoolyard bully or a racist is bad for Sam. *Qua* person or moral agent it is bad for Sam to be a bully or a racist. Each places an indelible black mark on Sam's character. The flourishing human being, or *eudaimon*, does not bully. Nor is she racist or homophobic. In light of these considerations, surely it is bad for Sam to be racist, a bully, or a homophobe. I think that all of this is correct. There is a deeply important sense in which Sam's being a bully is *bad for Sam*. However, this notion of "bad for" looks very different than the notion of "good for" and "bad for" that provided predictive power. The notion of "good for" and "bad for" we were originally working with applied not only to moral agents, but to plants and non-person animals as well. Lacking a convincing argument that the two usages are the same, moving to this latter notion of "bad for" would constitute a subtle bait-and-switch. I am not in possession of an argument that claims to show that the two uses of "good for" and "bad for" are the same—until I have, this last move is off limits.

I will conclude this section by highlighting an important similarity between the objection I pushed against a normative ethics based on role properties and the objection I pushed against a normative ethics based on the properties *being good for* and *being bad for*. My criticism of the role property reconstruction was founded on the observation that it can be bad to fulfill some role properties well. My criticism of the good for/bad for reconstruction was founded on the observation that it can be bad *simpliciter* for x to have the property of *being good for so-and-so*. In each case, we needed to consider a moral property in order to know if the non-natural normative property in question was worth attempting to bring about.

These criticisms are closely related to Moore's "open question argument" (Moore). While I think that Moore's open question argument has been shown to be unsound (Fumerton), it is nonetheless instructive. In light of the theory isomorphism defense of moral "realism" I embraced at the end of the previous chapter (and the failure of the argument schema I developed in the fourth chapter to show that comprehending application of moral predicates co-varies with the presence of some non-natural properties), the challenge for the "realist" is to show that the comprehending application conditions of some non-moral predicate are very nearly the same as the comprehending application conditions of some moral predicate. The failure of Moore's argument suggests that no successful general argument will demonstrate that the comprehending application conditions of all moral predicates are significantly distinct from the comprehending application conditions of all non-moral predicates. Nonetheless, the intuitive draw of Moore's open question argument demonstrates that our use of moral language appears to be significantly different from our use of non-moral language. There is a significant *prima facie* presumption against any attempt to construct a theory isomorphic to a moral theory out of non-moral predicates.

I do not take myself to have shown that one cannot construct an isomorph to moral theory out of the properties *good for* and *bad for*. I have, however, shown that any such attempt faces significant problems. I encourage others to explore more fully this strategy for defending ethical non-naturalism. I am not, however, convinced that it will be successful. I will, in hopes of finding an approach with greater promise, proceed to consider one final strategy for defending ethical non-naturalism.

Final Attempt: Vindicating the Moral Faculty

I have now offered four different attempts to apply the argument schema I developed in the fourth chapter in defense of ethical non-naturalism. The first attempt looked at the prospects of applying the argument schema to the primary moral predicates: "is good" and "is bad" *and* "is right" and "is wrong." Having given some reason to think that this first attempt would be unsuccessful, I considered the prospects for a defense of ethical non-naturalism rooted in virtue predicates. The next three attempts I considered were all, in one way or another, neo-Aristotelian. Each attempt offered to construct an isomorph to moral theory out of some predicate that is not generally taken to refer to a central moral property. I argued that each such attempt faces significant challenges. This suggests that the ethical non-naturalist ought to look for some new strategy. In the final sections of this chapter I will change gears entirely. Instead of attempting to offer a direct vindication of ethical non-naturalism by constructing an isomorph to moral theory out of predicates that can successfully instantiate the argument schema I developed in chapter four, I will instead opt for an indirect route. In short, I will argue that we have good reason to believe that the function of our moral faculty is to track non-natural normative properties and, in virtue of this fact, we have good reason to believe that there are non-natural *moral* properties.

Start by considering the following hypothesis:

[H] *The function of the mechanisms that ground our dispositions to make moral judgments is to recognize instantiations of non-natural normative properties*

My aim in this section is two-fold. First, I will argue that, if [H] is true, then we can vindicate ethical non-naturalism. Second, I will attempt to provide some reasons for thinking that [H] is true.

One might have an immediate concern regarding the plausibility of [H]: talk of our moral faculty having a function may suggest commitment to some kind of intelligent design theory. If one considers the paradigm instance of objects that have functions—a spell checker, a bottle opener, a chair, etc.—each has its function in virtue of having a designer. A methodologically naturalist defense of ethical non-naturalism had better not rely on the claim that humans have been designed. For [H] to be plausible, we need to have some other notion of function in mind. Luckily, it is widely accepted that evolution can ground function talk (Bekoff and Allen 254). We can be committed to the claim that *the function of the heart is to pump blood* without being committed to some version of a design theory.

For the time being, let us assume that [H] is true. How might the truth of [H] speak in favor of ethical non-naturalism? I hope that the following claim is uncontroversial: our moral judgments are non-accidentally related to some set of properties. Were this not the case, we would expect our moral judgments to be random. There would be no commonalities amongst the things that lead us to make various moral judgments. But we do see such commonalities. We tend to judge that causing harm is morally impermissible and that going out of one's way to help another is morally laudable. We can think of the point this way: were our moral judgments only accidentally related to some set of properties, normative theorizing would be impossible. The goal of normative ethics is to arrive at some set of principles that unifies our considered moral judgments. The identification of such a set constitutes the identification of some set of properties our moral judgments are non-accidentally related to. The fact that the project of arriving at a normative ethical theory has seemed worth pursuing is very good evidence that our moral judgments are non-accidentally related to some set of properties.

We can now pose this question: what set of properties are our moral judgments non-accidentally related to? [H] offers an answer to this question:

> [H] *The function of the mechanisms that ground our dispositions to make moral judgments is to recognize instantiations of non-natural normative properties*

If [H] is true, then our moral judgments are non-accidentally related to mind-independent non-natural normative properties. We can now present a dilemma to the non-naturalist's opponents.

Assume the truth of [H]. Consider the broad strokes strategy that composed the majority of this chapter. After "is good" and "is right" failed to fulfill the four conditions established in chapter four, the non-naturalist aimed to ground an isomorph to moral theory in various mind-independent non-natural properties that are not generally taken to be central to morality, namely the properties that co-vary with the comprehending application of various virtue ethical predicates. On one horn of the dilemma, the ethical non-naturalist is successful in this project. On the other horn of the dilemma, she is not. Suppose that the ethical non-naturalist is successful in grounding a theory isomorphic to moral theory in mind-independent non-natural properties. As I understand the goals of the ethical non-naturalism, on the first horn of the dilemma, the ethical non-naturalist has won. I am, however, not entirely optimistic about the prospects for this project. At the very least, I have been incapable of identifying a set of mind-independent non-natural properties that can be used in the construction of a theory that would vindicate the important commitments of the ethical non-naturalist. This pushes us towards the second horn of the dilemma: suppose that a theory isomorphic to moral theory cannot be constructed out of mind-independent non-natural properties that are not normally taken to be central to morality.

The plausibility of a given normative ethical account is a consequence of its ability to capture our considered moral judgments. If we cannot construct a moral theory out of some set of properties, P, then P cannot be the set of properties that our moral judgments are primarily responsive to. On the second horn of the dilemma, let "P" pick out any given set of mind-independent non-natural properties such that these mind-independent non-natural properties constitute a candidate for a set of properties that can ground an isomorph to moral theory *and* these properties are not generally taken to be central to morality, i.e. *being good* and *being right* are not members of P. Acceptance of the second horn of the dilemma forces us to reject the plausibility of a normative ethics constructed out of any given P. It further follows that we are forced to think that our moral faculty is responsive to some set of properties not included in P. Suppose that [H] is true: *the function of the mechanisms that ground our dispositions to make moral judgments is to recognize instantiations of non-natural normative properties*. If [H] is true then whatever properties are tracked by our moral faculty but fall outside of P are both mind-independent and non-natural. Furthermore, by hypothesis, we know that the properties in question are not properties that are generally considered not to be central to morality. What's left? Non-natural properties that are generally considered central to morality. Which properties are these? The right and the good! If we have good reason to believe that the deliverances of our moral faculty are the consequence of a non-accidental relationship between our moral faculty and mind-independent non-natural normative properties, the second horn of the dilemma forces us to posit the existence of non-natural *moral* properties in order to account for our moral judgments. On the second horn of the dilemma, the ethical non-naturalist gets the most straightforward vindication of her view available: we have good reason to believe that either the right or the good is a mind-independent non-natural property.

If we have good reason to believe [H] then the above dilemma gives us good reason to think that ethical non-naturalism can be vindicated. Do we have any reason to believe [H]? Despite opting for an indirect defense of non-naturalism, the argument schema I developed in chapter four will, nonetheless, play an important role. So far, I have argued that the following two types of properties are mind-independent and non-natural: role properties and the properties of *being good for some organism* and *being bad for some organism*. Each kind of property appears to play an important role in our moral judgments.

Consider how role properties influence our moral judgments. Imagine a teacher who opts to play video games instead of preparing lesson plans for class. We think ill of any such teacher. She has an obligation to her students and, by not preparing for class, she is failing to fulfill that obligation. Alternatively, consider a janitor who plays video games instead of preparing lesson plans. Here we do not think that the janitor has done anything wrong. The janitor is not a teacher. It would be odd were she to prepare lesson plans. What undergirds this difference in judgment regarding the teacher and the janitor? The answer appears to be: role properties. In not preparing a lesson plan the teacher fails to excel at fulfilling her role property. The same cannot be said of the janitor. Such examples can be produced *ad infinitum*, e.g. we judge a father more harshly for letting harm befall his child than we would if he let the same harm befall a stranger's child, we judge a fireman more harshly for not running into a burning building to save a child than we would any given civilian, we judge a professional truck driver more harshly for being bad at driving than we judge someone who does not drive for a living, etc. It appears that our moral judgments are non-accidentally related to role properties. The extent to which someone excels at their role appears to have an important influence on our moral judgments regarding that individual.

The same can be said about moral judgments and the properties *being good for* and *being bad for*. In the next chapter I will argue that disgust is a response to the property of *being bad for humans*. I will also point to evidence that disgust influences our moral judgments. It follows that the *good for/bad for* properties influence our moral judgments.

I have previously argued that role properties and the properties *being good for* and *being bad for* some organism are mind-independent and non-natural. Every instance of these properties influencing our moral judgments constitutes confirmation of [H]. Importantly, instances in which our moral faculty reliably tracks one of these two non-natural normative properties are ubiquitous. It appears that cases that confirm [H] are not hard to find. If [H] is well confirmed then we have good reason to believe that ethical non-naturalism is true.

An Objection Reconsidered

In the latter half of the previous chapter I raised an objection to the argument schema I developed there. The worry was as follows: demonstrating that comprehending application of a predicate improves predictive power may be sufficient to show that the predicate co-varies with some property; however, it is *not* sufficient to show that the co-varying properties we have identified are the properties that the predicate refers to. The argument schema I developed in the fourth chapter may give us reason to believe that some predicate co-varies with a mind-independent non-natural property; however, it does not give any reason to believe that said predicate *refers to* a mind-independent non-natural property. Consequently, the argument schema cannot demonstrate that there are *moral* properties, which is putatively what the realist needs to show.

At the conclusion of the fourth chapter I argued that the ethical non-naturalist should not feel compelled to demonstrate that there are non-natural *moral* properties. If one admits that there are no non-natural moral properties, one cannot be properly called an "ethical non-naturalist"; however, this is largely a taxonomical issue. Demonstrating that there are mind-independent non-natural properties that fit the description of moral properties is enough to vindicate the important commitments of ethical non-naturalism. If one can show that a theory isomorphic to moral theory describes an ontologically robust aspect of our universe, it is difficult to see why one ought to be worried about the accuracy of calling this theory a "moral theory."[34]

[34] This claim is not entirely accurate. It is important to keep in mind the caveat I developed in the previous chapter regarding disjunctive properties.

Nonetheless, after arguing that the ethical non-naturalist can be content with a theory that is not, in fact, a version of ethical non-naturalism, I promised to provide a more conciliatory response. I hope that, now, I am in a position to do so. My considered argument for ethical non-naturalism concluded with a dilemma. I will close this chapter by considering the force of the objection with regard to the second horn of the dilemma.

On the second horn of the dilemma, a theory isomorphic to a moral theory cannot be constructed out of mind-independent non-natural properties that are not taken to be central to morality. That is, on the second horn of the dilemma, were we to take the entire set of normative mind-independent non-natural properties, excluding the putative properties of *being right/wrong* and *being good/bad*, we would be unable to construct a theory isomorphic to moral theory out of this set of properties. Call this set of mind-independent non-natural properties, excluding the putative properties *being right* and *being good*, "P." The aim of normative theorizing is to offer a theory that captures the majority of our considered moral judgments. If we cannot arrive at a theory isomorphic to moral theory based on P, it must be the case that some of our considered moral judgments are non-accidentally related to a property that does not fall within P. But [H] is well confirmed. We have reason to believe that our moral faculty reliably tracks mind-independent non-natural normative properties. That is, we have reason to believe that whatever property our moral faculty tracks that is outside of P, that property is normative, mind-independent, and non-natural. What kind of normative, mind-independent, and non-natural property falls outside of P? By hypothesis, there are only two candidates for normative, mind-independent, non-natural properties that fall outside of P: *being good/bad* and *being right/wrong*.

On the second horn of the dilemma, the worry that I have failed to identify *moral* properties has no bite. The structure of the second horn of the dilemma assures that the normative, mind-independent, non-natural properties in question are moral properties. Even though I am unconvinced that the ethical non-naturalist should feel pressure to meet the burden of demonstrating that moral terms refer to mind-independent non-natural properties, if we end up embracing the second horn of the dilemma, the ethical non-naturalist will, nonetheless, have met this burden.

Conclusion

This chapter concludes my primary argument for ethical non-naturalism. The argument relies on *a posteriori* considerations regarding the limits of computational power in the universe and the extent to which the hypothesis [H] is well confirmed. The considered argument I have offered in favor of ethical non-naturalism ought to be acceptable from the perspective of the methodological naturalist. If the argument is successful, I have shown that, if one is committed to a scientific worldview, one is thereby committed to ethical non-naturalism.

I will, however, conclude this section on a less-than-triumphant note. I am not entirely happy with the argument I have offered. My central concern is that [H] is only one of many competing hypotheses. Though we have some evidence for [H], the evidence is far from overwhelming. Thus, though the fact that there is some confirmation for [H] gives us some reason to believe in ethical non-naturalism, the case is far from closed. It may be that, though our moral faculty reliably recognizes mind-independent non-natural normative properties, it also responds to some range of more mundane properties. The evidence I have offered gives us some reason to think that, if our moral faculty tracks some property, that property is both mind-independent and non-natural. In turn, this gives us some reason to prefer ethical non-naturalism. The case is not, however, as strong as I would like it to be. There is plenty of room for my opponent to push back. In the final two chapters I will attempt to strengthen the case I have offered so far by sketching an etiological account of our moral faculty. It is my hope to show that [H] can help explain why possessing a moral faculty would be evolutionarily advantageous. If my project in the final two chapters is successful, I will have accomplished two distinct goals. First, I will have offered reasons, largely independent of the content of this chapter, to accept ethical non-naturalist. Second, I will have given us further reason to accept [H] by showing that [H] is consilient.

CHAPTER SIX:
THE EVOLUTION OF MORALITY

Introduction

It is best to open this chapter with a brief summary of where we have been. In the first chapter I sketched the meta-ethical territory and familiarized the reader with the meta-ethical view I intend to defend: ethical non-naturalism. In the second chapter I offered some reason to think that the moral realist has the correct account of moral semantics. The third chapter provided a very brief summary of what would constitute a methodologically naturalist defense of ethical non-naturalism. The true work of the dissertation started in the fourth chapter. There I proposed a method for identifying non-natural properties. In the fifth chapter I considered a variety of ways in which one might attempt to apply this method in defense of ethical non-naturalism. The results were mixed. Application of the method appeared to reveal that role properties and the properties of being good for/being bad for were both non-natural. Unfortunately, there is no obvious way to construct a moral theory out of either type of property. Instead, I argued that, if we have good reason to believe that the function of our moral faculty is to recognize non-natural normative properties, we have good reason to accept ethical non-naturalism. I concluded by offering some reason to think that the function of our moral faculty is to recognize non-natural normative properties. As I noted at the end of the last chapter, I am not entirely happy with the direction the defense has taken.

The crux of the problem is that, for the defense to be successful, we must have good reason to believe that our moral faculty is responsive to non-natural normative properties. Demonstrating that this is the case is no small task. While I am in no position to offer a decisive defense of this account of our moral faculty, I think that there are reasons to prefer the hypothesis. In this chapter I will consider how the hypothesis that *the function of the mechanisms that ground our dispositions to make moral judgments is to recognize instantiations of non-natural normative properties* relates to evolutionary considerations.

I have two primary aims. First, I want to build on the defense of ethical non-naturalism I have thus far offered and provide a response on behalf of the non-naturalist to a current strain of anti-realist argument that purports to show that a commitment to methodological naturalism and a commitment to ethical non-naturalism are incompatible. Some anti-realists, Sharon Street and Richard Joyce in particular, have argued that a commitment to ethical non-naturalism is incompatible with the acceptance of a neo-Darwinian account of evolution. Given that neo-Darwinian evolutionary theory appears to be one of our most highly confirmed scientific theories, it would be very bad news for a methodologically naturalist defense of ethical non-naturalism were the two incompatible. In the first half of this chapter I will offer a putative account of our evolutionary history intended to undermine the anti-realist's objection.

In the second half of the chapter, I want turn the anti-realist's argument on its head. In light of the putative evolutionary history I have offered, I will argue that the ethical non-naturalist is in a position to offer the best explanation of the etiology of the moral faculty. The second half of this chapter is intended to significantly supplement the arguments for ethical non-naturalism I have thus far developed. Scientific theories are taken to be particularly well confirmed when they are positioned to explain disparate phenomena. Chapters four and five were dedicated to offering an explanation of the gains in predictive power we get when we accurately apply certain predicates. It would be a significant boon to the account I developed in those chapters were the explanation I offered there easily expandable to offer an explanation of an unrelated state-of-affairs, namely, the etiology of our moral faculty. It is my hope that this chapter will not only defuse an important anti-realist criticism but will offer an independent line of evidence in favor of the thesis that the function of whatever grounds our dispositions to make moral judgments is to recognize non-natural normative properties.

Street's evolutionary debunking argument

Two philosophers in particular have argued that moral realism—specifically ethical non-naturalism—is incompatible with a neo-Darwinian account of the etiology of the human species. Superficially, the argument offered by Sharon Street is distinct from the argument offered by Richard Joyce. I am of the opinion that this apparent difference is just that: superficial. It would, however, require a significant amount of dialectical disambiguation to demonstrate that this is the case. As such, I will forego a careful examination of the similarities between the two arguments. Instead, I will focus exclusively on Street's formulation. All that the reader need know is that both Street and Joyce think that all plausible accounts of the etiology of our moral faculty that are compatible with moral realism lead to moral skepticism. In responding to Street I will offer an evolutionary account intended to serve as a counter-example to this claim and, in doing so, simultaneously diffuse Joyce's worry.

Before offering a presentation and disambiguation of Street's argument, I first need to distinguish between two closely related views: *ethical non-naturalism* and *normative non-naturalism*. The reader is already intimately familiar with the first. The *ethical non-naturalist* is committed to thinking that (1) moral properties exist, (2) moral properties are mind-independent, and (3) instantiations of moral properties are not token-token identical with the instantiation of any set of microphysical properties. The *normative non-naturalist* is committed to an analogous set of claims, but about normative properties: (1) normative properties exist, (2) normative properties are mind-independent, and (3) instantiations of normative properties are not token-token identical with the instantiation of any set of microphysical properties.

What would constitute a "normative" property? While one can distinguish between *evaluative properties* and *normative properties*, I will not do so here. I will use the term "evaluative" and "normative" interchangeably. Evaluative/normative facts are "of the form that X is a … reason to Y, that one should or ought to X, that X is good, valuable, or worthwhile, that X is morally right or wrong, and so on" (Street 110). It would be a normative/evaluative fact were some state-of-affairs to *count in favor of x* or *justify x* (Street 156). Ethical non-naturalism is a subset of the views that count as versions of normative non-naturalism; moral facts are a type of normative fact. In some ways, my dissertation is anachronistic. Increasingly, philosophers are apt to argue about the ontological status of *normative facts* instead of couching the same debate in terms of *moral facts*.

Following this trend, Street's argument is aimed at normative realism, and in particular, normative non-naturalism. As ethical non-naturalism constitutes a subset of the views that count as versions of normative non-naturalism, if Street's argument is effective then I am wrong that a commitment to a scientific worldview entails commitment to ethical non-naturalism. Vindication of ethical non-naturalism, however, requires more than showing that Street's argument fails with regard to normative non-naturalism. I must further show that it fails with regard to ethical non-naturalism. This dual project is the task of a significant portion of what follows.[35]

[35] Before proceeding I need to make a note about an immediate problem for Street's argument. Street takes herself to be arguing against normative realism. As she, and we, understand(s) "normative," epistemic norms are included in the extension of the term. Thus, if Street's argument is successful, it applies to epistemic realism. But it is not implausible to think that any argument against epistemic realism is self-defeating. For if the argument is successful then epistemic realism is false; it is plausible to think that the truth of epistemic realism is a precondition of having any reason to believe anything. Put another way, epistemic norms provide the link between Street's premises and her conclusions. The denial of epistemic realism is tantamount to denying the existence of these norms, and thus tantamount to denying the claim that Street's premises give us reason to believe her conclusion. But if Street's argument is self-defeating, then it fails to constitute a challenge for moral realism and can safely be ignored (Machery and Mallon).

Street seems content with the expansion of her argument to epistemic norms (156)—indeed, Street's argument closely mirrors a pyrrohnian argument offered by Plantinga—however, I think that Street's argument can be presented in such a way that it challenges normative realism *exclusive* of epistemic norms. I do not intend to offer a formal presentation of a modified version of Street's argument; doing so would take us too far afield from the present topic. I can, however, gesture at my reasons for thinking that Street's argument need not be self-defeating.

Street wants to offer a general argument against all of normative realism; however, we should not take the putative scope of her argument at face value. It may apply better to some

Street starts with a presupposition that I do not intend to challenge: our dispositions to form evaluative judgments are a consequence of our evolutionary history.[36] It is plausible to think "human cognitive traits are (in some cases) just as

norms than to others. The general conclusion will only follow if her argument applies to realism about *all* kinds of normative claims. I take it that something like the following principle is both (1) a rational requirement and (2) suggests that if Street's argument is effective against robust realism about one kind of norm, it will be effective against robust realism about all norms:

> [Non-*ad hoc*] If an argument with premises, p_1 through p_n, successfully instantiated by propositions in domain D_1 and logical structure L, leads to conclusion C_1 in domain D_1, then for any D_z for which premises p_1 through p_n can be successfully instantiated, conclusion C_z in domain D_z follows.

Put more succinctly: you don't get to pick and choose the targets of your arguments. If an argument works regarding, e.g., pragmatic norms, then presuming one can re-create an analog of all of the original premises but this time with regard to epistemic norms, one must accept the same conclusion about epistemic norms as one accepts about pragmatic norms. Acceptance of [Non-*ad hoc*] gives some reason to think that if Street's argument works with regard to realism about moral or pragmatic norms it will also work with realism about epistemic norms. Nonetheless, I think that we can accept [Non-*ad hoc*] without being forced to think that Street's argument applies equally well to epistemic norms as it does to other types of norms.

The argument to this end is extremely simple. In short, if one attempts to instantiate Street's argument with regard to epistemic norms, one ends up with an argument of *different logical structure* than if one attempts to instantiate Street's argument with regard to non-epistemic norms. When the argument is applied to epistemic norms, the conclusion ranges over the relation between the premises and itself. When the argument is applied to non-epistemic norms, the conclusion does not range over the relation between the premises and itself. It follows that the logical relations between the premises and conclusion are different for attempts to instantiate the argument with regard to epistemic norms as compared to attempts to instantiate the argument with other types of norms. It follows that Street's argument can be successful with regard to all norms other than epistemic norms while not falling prey to self-defeat. For simplicities sake, in what follows I will use the terms "normative" and "evaluative" to pick out the subset of the normative/evaluative *exclusive* of epistemic norms.

[36] While I will be granting Street this premise, it should be noted that, given how Street appears to interpret the premise, her claim is quite controversial. Street appears to think that our dispositions to form moral judgments are the consequence of micro-selection. Put another way, Street appears to assume that our dispositions to form moral judgments are not the consequence of evolution selecting for a very general cognitive capacity, e.g. the ability to reason. It is open to the realist to hold that, just as our ability to do calculus is a consequence of our general ability to reason, our dispositions to think about things morally is also a consequence of our general ability to reason. Once we have evolved the general ability to reason, we can reason about whatever we would like, morality included. If there is an evolutionary puzzle with regard to our ability to reason, it certainly does not present a unique challenge for the moral realist.

susceptible to Darwinian explanation as human physical traits…" (Street 113). Furthermore, Street thinks that it is particularly plausible to suppose that our moral faculty is one of the cognitive traits susceptible to neo-Darwinian explanation:

> [N]ote the potentially phenomenal costs and benefits, as measured in the Darwinian currency of reproductive success, of accepting some evaluative judgements rather than others. It is clear, for instance, how fatal to reproductive success it would be to judge that the fact that something would endanger one's survival is a reason to do it, or that the fact that someone is kin is a reason to harm that individual… In contrast, it is clear how beneficial … it would be to judge that the fact that something would promote one's survival is a reason in favor of it, or that the fact that something would assist one's offspring is a reason to do it. (Street 144)

It should be noted that the assumption that our dispositions to make moral judgments are the result of evolution, i.e. "moral nativism", is a topic of some controversy. Like Street, I will take some version of weak moral nativism for granted. Nonetheless, we ought not pretend to have better reason to believe moral nativism than is in fact the case. The strategy in this chapter will be familiar. My considered arguments are conditional on some set of empirical assumptions. It may be the case that some, or

In the remainder of the dissertation, I will assume that our ability to make moral judgments is not a consequence of our more general ability to reason. Acceptance of this premise constitutes a *significant* concession to the moral anti-realist. Historically, moral realists have held that our ability to make veridical moral judgments was a consequence of our ability to reason. Rejecting this claim about the relationship between reason and moral judgment is tantamount to abandoning the historically most successful strategy for defending moral realism.

I have two reasons for granting Street this premise. First, the project of this dissertation is to see how far we can get in attempting to offer a defense of moral realism on ground that is generally considered to be favorable to the anti-realist. Acceptance of Street's etiological premise is part-and-parcel of this project. Second, in light of recent results from cognitive science, it is not clear to me that it is still tenable to hold onto the analogy between mathematical judgments and moral judgments. Many contemporary cognitive scientific models hold that emotions play a fundamental role in the formation of moral judgments (for a more in depth discussion, see chapter two). In so far as these models are accurate, it would appear that it is no longer sufficient to merely say that the cognitive mechanisms responsible for our moral judgments are the same as the cognitive mechanisms responsible for our more general ability to reason.

all, of these empirical assumptions turn out to be false. Until scientific consensus is reached, the best I can do is to base my arguments on well-supported conjecture.

Having assumed the truth of some version of weak moral nativism, Street proceeds to present a dilemma aimed at the realist about evaluative truth. Street asks the following question: what influence did evaluative truth have on the evolution of our moral faculty? The realist has, broadly speaking, two options available to her. She can either hold that there is *no* relationship between evaluative truth and the evolution of our moral faculty **or** she can hold that there *is* some relationship between evaluative truth and the evolution of our moral faculty. Denial of a relationship between evaluative truth and the evolution of our moral faculty constitutes the first horn of the dilemma. Asserting such a relationship constitutes the second horn of the dilemma.

Imagine that the realist denies that there is a relationship between evaluative truth and the evolution of our moral faculty. Street argues that, if this is the case, we have good reason to doubt the veracity of any of our evaluative judgments. She offers the following analogy to make her point:

> The key point to see about this option is that if one takes it, then the forces of natural selection must be viewed as a purely distorting influence on our evaluative judgements, having pushed us in evaluative directions that have nothing whatsoever to do with the evaluative truth. On this view, allowing our evaluative judgements to be shaped by evolutionary influences is analogous to setting out for Bermuda and letting the course of your boat be determined by the wind and tides: just as the push of the wind and tides on your boat has nothing to do with where you want to go, so the historical push of natural selection on the content of our evaluative judgements has nothing to do with evaluative truth… If we take this point and combine it with the first premise that our evaluative judgements have been tremendously shaped by Darwinian influence, then we are left with the implausible skeptical conclusion that our evaluative judgments are in all likelihood mostly off track, for our system of evaluative judgments is revealed to be utterly saturated and contaminated with illegitimate influence. (Street 121-122)

If evolutionary pressures importantly influenced the content of our moral judgments and there is no relationship between evolutionary pressures and the evaluative truth, it would be a miracle were our moral judgments to track evaluative truth.[37]

Let us set aside the first horn of Street's dilemma. I am interested in the prospects of the ethical non-naturalist successfully grasping the second horn of the dilemma. On the second horn, the normative realist is committed to thinking that evolution pushed our moral faculty in direction of reliable judgments about normative truth. On the second horn of the dilemma, Street argues that the realist owes us some account of what this relationship between normative truth and fitness conducivity is supposed to be.

Street nicely summarizes a broad strokes sketch of the structure she thinks any such account must take:

> According to [the realist's] hypothesis, our ability to recognize evaluative truths, like the cheetah's speed and the giraffe's long neck, conferred upon us certain advantages that helped us flourish and reproduce. Thus, the forces of natural selection that influenced the shape of so many of our evaluative judgments need not and should not be viewed as distorting or illegitimate at all. For the evaluative judgments that it proved most selectively advantageous to make are, in general, precisely those evaluative judgments which are true. (Street 126)

[37] I have argued, elsewhere, that the first horn of Street's dilemma is subtly question begging. I offered the following counter-analogy. Imagine that, after jumping into one's boat without any navigational gear or map, one eventually washes up on a beach that sports a thirty-foot tall billboard emblazoned with the words "Welcome to Bermuda!" I take it that, were we to find ourselves in such circumstances, we'd have *very good* reason to believe that we'd ended up in Bermuda (Graber). A similar point can be made regarding evolution and evaluative truths. If we have independent reason to believe that our evaluative judgments track evaluative truth then, whether or not there is a relationship between evaluative truth and evolutionary pressures, we have good reason to believe that—miracle or otherwise—evolutionary pressures ended up pushing in the direction of evaluative truth. In the fourth and fifth chapters of this dissertation, I argued that we have good reason to believe that our evaluative judgments track evaluative truth. If those arguments are successful then the normative realist can, without worry, grasp the first horn of Street's dilemma.

Street labels any etiological sketch of this structure a *tracking account*. She contrasts the tracking account with the *adaptive link account* whereby,

> tendencies to make certain kinds of evaluative judgments rather than others contributed to our ancestors' reproductive success not because they constituted perceptions of independent evaluative truths, but rather because they forged adaptive links between our ancestors' circumstances and their responses to those circumstances, getting them to act, feel, and believe in ways that turned out to be reproductively advantageous. (Street 127)

For reasons that are rather opaque, Street thinks that the *tracking account* and the *adaptive link account* are mutually incompatible. Furthermore, Street also thinks that the realist is committed to giving some version of the *tracking account* whereas it is open to the anti-realist to give some version of the *adaptive link account*. As they are both candidates for an etiological explanation, we ought to prefer whichever explanation is best. If the adaptive link account provides a better explanation of the evolution of our moral faculty and only the anti-realist can offer an adaptive link account, then we ought to prefer evaluative anti-realism to evaluative realism. The evaluative anti-realist, but not the evaluative realist, is in a position to explain why our normative judgments are by-and-large true.

Central to Street's argument is the claim that the adaptive link account provides a better explanation of the etiology of our moral faculty than does the tracking account. I will not consider her arguments to this end because it seems clear to me that the adaptive link account and the tracking account are perfectly compatible. By way of arguing for their incompatibility Street writes:

> For illustration of the differences between the adaptive link account and the tracking account, consider a few examples. Consider, for instance, the judgement that the fact that something would promote one's survival is a reason to do it, the judgement that the fact that someone is kin is a reason to accord him or her special treatment, and the judgement that the fact that someone has harmed one is a reason to shun that person or retaliate. Both the adaptive link account and the tracking account explain the widespread human tendencies to make such judgements by saying that making them somehow contributed to reproductive success in the environment of our ancestors. According to the tracking account, however, making such evaluative judgements contributed to reproductive success because they are *true*, and it

> proved advantageous to grasp evaluative truths. According to the adaptive link account, on the other hand, making such judgements contributed to reproductive success not because they were true or false, *but rather because they got our ancestors to respond to their circumstances with behavior that itself promoted reproductive success…* [emphasis added] (Street 128-129)

Nothing Street says here gives us any reason to think that the two accounts are incompatible. The crux of the *adaptive link account* is the claim that our dispositions to make evaluative judgments was fitness conducive "because they forged adaptive links between our ancestors' circumstances and their responses to those circumstances, getting them to act, feel, and believe in ways that turned out to be reproductively advantageous" (ibid.). But this is entirely compatible with thinking that "the evaluative judgments that it proved most selectively advantageous to make are, in general, precisely those evaluative judgments which are true" (Street 126). Presumably, the realist thinks that making true evaluative judgments was fitness conducive *because* doing so led our ancestors to "act, feel, and believe in ways that turned out to be reproductively advantageous." It is difficult to imagine some mechanism not on this list that might have linked true evaluative judgments to fitness. In short, the normative realist is not committed to defending the *tracking account* **in face of** *the adaptive link account*. Much more reasonably, the normative realist must only give some account of how tracking the moral truth could reasonably have constituted an adaptive link.

This is not, however, to say that Street's argument need not worry the realist. Despite the compatibility of the tracking account and the adaptive link account, and subsequently the invalidity of Street's considered argument, Street's work suggests *two* independent explanatory challenges that the normative realist must meet. The first explanatory challenge is unique to the normative realist; both the normative realist and her opponents share the second explanatory challenge.

In an exemplary article, David Enoch helps to clarify the nature of the first explanatory challenge. Like Street, Enoch develops his argument via analogy. I quote him at length:

> Mathematicians are remarkably good when it comes to their mathematical beliefs. Almost always, when mathematicians believe a mathematical proposition p, it is indeed true that p, and when they disbelieve p (or at least when they believe not-p) it is indeed false that p. There is, in other words, a striking correlation between mathematicians' mathematical beliefs… and the mathematical truths. Such a striking correlation calls for explanation. But it doesn't seem that mathematical Platonists are in a position to offer any such explanation. The mathematical objects they believe in are abstract, and so causally inert, and so they cannot be causally responsible for mathematicians' beliefs; the mathematical truths Platonists believe in are supposed to be independent of mathematicians and their beliefs, and so mathematicians' beliefs aren't causally (or constitutively) responsible for the mathematical truths. Nor does there seem to be some third factor that is causally responsible for both. What we have here, then, is a striking correlation between two factors that Platonists cannot explain in any of the standard ways of explaining such a correlation—by invoking a causal (or constitutive) connection from the first factor to the second, or from the second to the first, or form [sic] some third factor to both. But without such an explanation, the striking correlation may just be too implausible to believe, and… so is mathematical Platonism. ("The epistemological challenge to metanormative realism" 9)

The crux of Enoch's reconstruction of Street's argument is the following observation: if it is highly implausible to think that a correlation is brute, the correlation requires explanation. If no explanation can be given, we would do well to give up whatever theoretical commitments require us to posit the correlation in question.

Enoch thinks that the robust realist—a person who holds a position closely related to my understanding of "ethical non-naturalism"—finds herself in a position analogous to the mathematical Platonist. There is a striking correlation between our moral judgments and moral truth.[38] Enoch takes the robust realist to be committed to

[38] More needs to be said to precisify the correlation Enoch has in mind. If you consider the amazing variety of day-to-day moral judgments (e.g. responses to the question, "is gay marriage morally permissible?), it is not clear that there is much of a correlation between these judgments and the moral truth. Enoch appears to have in mind judgments, not about *ultima facie* permissibility, but rather moral judgments that involve neither the weighing of competing moral properties nor empirical considerations: pain is intrinsically bad, we have a *prima facie* obligation to keep our promises, etc.

the claim that normative properties do not have causal powers. Furthermore, Enoch thinks that positing some sort of quasi-perceptual access to normative properties cannot solve the normative realist's problem:

> [E]ither this quasi-perceptual faculty is causal (like perception), putting us in causal relations with the normative truths, or it isn't. If it is, then the normative truths cannot be causally inert, as on Robust Realism they must be. And if this faculty is not causal, then for everything thus far said it is very hard to see how it can help in explaining the correlation that needs explaining. ("The epistemological challenge to metanormative realism" 12)

I think that Enoch is wrong to think that the problem for normative realism lies in the correlation between normative judgments and normative truth. For two reasons, I do not think there is any real trouble for the realist here. First, unlike Enoch, I am not committed to the claim that non-natural normative properties lack causal powers. Thus, unlike Enoch, I lack any in principle reason to hold that we cannot be causally related to moral properties. Second, I have difficulty understanding why Enoch thinks that a non-causal quasi-perceptual faculty cannot solve the problem. Reliance on the quasi-perceptual relationship of direct acquaintance has a long and storied history in epistemology. Plenty of epistemologists have held that direct acquaintance with mathematical and logical facts explains our knowledge of math and logic. If one wants to show that a quasi-perceptual faculty is insufficient to explain the correlation between, for example, our mathematical beliefs and the mathematical truth, much more needs to be said.

There is, however, another surprising correlation that the normative realist must explain. On the second horn of Street's dilemma, the normative realist is committed to the view that tracking evaluative properties, by way of making true evaluative judgments, is fitness conducive. The most obvious, and perhaps only, explanation of this correlation is that making true evaluative judgments gets us to "act, feel, and believe in ways that turned out to be reproductively advantageous" (Street 129). Instead of solving the problem, offering this putative explanation only pushes the worry up a level. Now the question is: why is acting, feeling, and believing in line with evaluative truths fitness conducive?

The normative realist is committed to thinking that normative properties are, in an important sense, independent of humans. But if normative properties are, in an important sense, ontologically independent from us, what could possibly explain the fact the acting, feeling, and believing in line with normative truth was fitness conducive? At least on face, these two sets of properties appear to be independent: evaluative properties *and* the property of *being fitness conducive*. Unless the normative realist can give some reason to think that there is an important connection between the two types of properties, she is committed to a very implausible brute correlation. Explaining the correlation between *being a mind-independent evaluative property* and *being fitness conducive to track* constitutes the normative realist's first explanatory burden.

Both the normative realist and her opponent share the second explanatory burden. We have presupposed some version of weak moral nativism. That is, we have presupposed that it was evolutionarily advantageous to make evaluative judgments very much like the evaluative judgments we are disposed to make. This fact requires explanation. What is it about the evaluative judgments we are disposed to make that lead them to be fitness conducive? Any answer to this question will require the identification of some set of external world properties that our evaluative judgments are responsive to. The features of the world that determine fitness are found in the external world. Cliffs, predators, illness, starvation, etc. determine whether any given individual will pass on her genes. Thus, if our evaluative judgments do not track properties of the external world, it would be a miracle were making various evaluative judgments fitness conducive. If one wants to explain the fact *that it is fitness conducive to make the kind of evaluative judgments we are disposed to make*, one must posit the existence of some set of external world properties that our evaluative judgments track (Graber 595).

It is important to see that, for the realist, these two explanatory challenges are closely related. The realist has a ready answer to the question, "what set of external world properties do our evaluative judgments track?" The answer is: our evaluative judgments track *evaluative properties*. Of course, claiming that our evaluative judgments are responsive to evaluative properties can only explain why making evaluative judgments is fitness conducive if there is an important relationship between *being an evaluative property* and *being fitness conducive to be responsive to*. Being able to give some reason to think there is an intimate relationship between evaluative properties and fitness conducivity is exactly what is required in order to fulfill the other explanatory burden.

The evolutionary debunking argument: a partial answer

A pre-established harmony?

One of the more perplexing problems in philosophy involves determining the specifics of the relation, if any, between our mental states and our bodily states. Leibniz offered a unique account of this relationship. Leibniz held that our mental states and bodily states were metaphysically distinct; there was no causal or constitutive relationship between the two. But there is a striking correlation between our mental states and our bodily states, e.g. when my finger gets pricked by a pin, I feel pain. This correlation demands explanation. Leibniz famously held that there was a *pre-established harmony* between our mental states and our physical states: "created minds and bodies are programmed at creation such that all their natural states and actions are carried out in mutual coordination" (Kulstad and Carlin). As Leibniz understood the universe, G-d had created the universe such that there was a pre-established harmony between our minds and bodies.

Most realist attempts to respond to the two explanatory burdens I presented take the form of arguing for pre-established harmonies. As befits the tone of contemporary philosophy, G-d is left out of the picture. Instead, philosophers argue that there are necessary relations between evaluative facts and evolutionary pressures—a pre-established harmony. By way of illustration, I will consider three responses to Street that fit the pre-established harmony mold.

David Enoch responds to Street's dilemma by arguing for a pre-established harmony between *goodness* and *survival* or *reproductive success*. Enoch plausibly assumes that "survival or reproductive success… is at least somewhat good. Not, of course, that it is always good, or that its positive value is never outweighed by other considerations…" ("The epistemological challenge to metanormative realism" 18). If we grant this assumption, we find that there is a pre-established harmony between normative truth and evolutionary pressures that lets us explain how selective pressures might have pushed in the direction of true normative judgments:

> Survival … is good; so behaving in ways that promote it is (pro tanto) good; but one efficient way of pushing us in the direction of acting in those ways is by pushing us to believe that it is good to act in those ways. And in fact, as we have just seen, it *is* good so to act. So the normative beliefs this mechanism pushes us to have will tend to be true. ("The epistemological challenge to metanormative realism" 19)

Enoch starts by assuming a normative truth: survival is good. He then argues that, if we grant that survival is good, we would expect evolution to (1) push us in the direction of believing that it is good to act in ways that are conducive to survival and (2) in virtue of the original normative assumption, these beliefs will tend to be true.

Erick Wielenberg offers an alternative response to Street. Wielenberg relies on our notion of rights to argue for a pre-established harmony between evolutionary pressures and normative truth. He writes:

> Despite various cultural differences, human beings normally believe that there are certain things that others simply ought not do to them, for example, rape them, enslave them, steal from them, or kill them for entertainment. It is not hard to see how the disposition to form such beliefs might be fitness-enhancing… Viewing ourselves as possessing boundaries that may not be transgressed no matter what provides a distinctive kind of motivation to resist such transgressions by others. Holding such beliefs disposes one to resist behavior on the part

> of others that typically dramatically decreases one's prospects for survival and reproduction. (Wielenberg 445)

Wielenberg further notes that,

> While there are various theories about the foundation of rights, it is widely agreed that if rights exist at all, their presence is guaranteed by the presence of certain cognitive faculties. The cognitive faculties in question are either the very ones required to form beliefs about rights or are closely linked to such faculties. If you think you possess moral barriers, then you do… (Wielenberg 449)

Wielenberg's pre-established harmony argument has two parts. In the first, Wielenberg demonstrates that it is fitness conducive to believe that one has rights. In the second half of his argument, Wielenberg argues that, if one believes that one has rights, then one has rights. Thus, if evolution successfully pushes us in the direction of having a disposition to believe that we have rights, *we in fact have rights*. According to Wielenberg's preferred account, the surprising correlation between true normative judgments and selective advantage is explained by a necessary truth about the pre-conditions for possessing rights.

Knut Olav Skarsaune takes a third approach to demonstrating a pre-established harmony between selective pressures and true normative judgments. Skarsaune argues that pain, or any relevantly similar mental state, is intrinsically bad for the agent that experiences it. He further argues that it is plausible to think that pain, or some relevantly similar mental state, is an evolutionary requirement: it motivates organisms to avoid harmful stimulus. Furthermore, essential to pain's ability to motivate organisms to avoid painful stimulus is the fact that, from an organism's perspective, pain is *to be avoided*. Thus, there is a pre-established harmony between selective pressures and true normative judgments. The nature of pain is such that, necessarily, the judgment (or proto-judgment) that *pain is to be avoided* is true. Furthermore, some pain-like mental state may be an evolutionary requirement and a partial requirement of a pain-like mental state conferring selective advantage is that we judge (or proto-judge) that the pain-like mental state is *to be avoided* (Skarsaune).

In a moment, I will offer my own response to the dual explanatory challenge raised by Street's dilemma. In light of the three pre-established harmony arguments I have just presented, it may be difficult to see why I feel the need to offer a novel version of a familiar argumentative strategy. I earlier argued that Street's dilemma highlighted two distinct explanatory challenges. The first challenge is shared by everyone: some account is owed of the selective advantage conferred by our dispositions to make evaluative judgments. The second challenge is unique to the normative realist: some explanation is owed of the surprising coincidence between normative truth and selective advantage. Note that each of the above pre-established harmony arguments fares quite well as a response to both explanatory challenges, but only for a very tightly constrained set of normative judgments/truth. The pre-established harmony accounts offered by both Enoch and Skarsaune fall immediate prey to a challenge familiar from the fifth chapter of the dissertation: it is unclear how one could construct an isomorph to moral theory from, in Enoch's case, the goodness of reproductive fitness, or in Skarsaune's case, the fact that pain is *bad for* the organism that experiences it. Enoch's and Skarsaune's arguments are enough to vindicate *normative realism* from the dual explanatory challenge raised by Street; however, they will not do to vindicate *moral realism*.

The case is marginally more complicated when it comes to Wielenberg's approach. Wielenberg draws heavily on the notion of a right. *Having a right to such-and-such* is certainly a moral property. Wielenberg's attempt to demonstrate a pre-established harmony, if successful, is sufficient for a partial vindication of moral realism from the dual explanatory challenges. Nonetheless, I take it that the vindication is *only* partial. The trouble is this: it is not clear that many of our most dearly held moral judgments can be explicated in terms of rights talk. We tend to make a broad range of moral judgments that appear to be driven by considerations of virtue or considerations of maximization of goodness. Often, we think that *doing the virtuous thing* or *maximizing goodness* is morally obligatory even though doing so may require the violation of a right. Wielenberg's pre-established harmony account is sufficient to demonstrate that some subset of our moral judgments can fulfill the dual explanatory burden. As such, Wielenberg's pre-established harmony account is enough to vindicate moral realism. Nonetheless, we can do better. Acceptance of Wielenberg's account still leaves the moral realist with a significant explanatory burden: presuming that all of the evaluative judgments we are disposed to make cannot be re-expressed in terms of judgments about rights, some explanation is owed of the selective advantage offered by those evaluative judgments we make that are not judgments about rights. If selective pressures are responsible for our dispositions to make evaluative judgments then presumably our dispositions to judge that *it is morally obligatory to maximize goodness* and that *it is morally obligatory to act virtuously* are the consequence of selective pressures. It is open to the moral realist to hold that these types of judgments are false. This allows the realist to escape the second explanatory burden; that is, the realist does not need to explain the correlation between these kind of normative truths and selective advantage. The realist does, however, still owe some evolutionary account of how we came to be disposed to make these kinds of judgments. I hope to offer an account capable of fulfilling this further explanatory burden. That is, I hope that the account I offer can not only explain the correlation between normative truth and selective advantage, but also explain, for a broad range of evaluative judgments, why evolution pushed us in the direction of making these judgments.

Reliable moral judgments: a spandrel

In the previous chapter I presented my considered argument for the existence of non-natural moral properties. At the time, I noted that the argument was less direct that I would have liked. In chapter four I developed an argument schema for identifying non-natural properties. Unfortunately, the argument schema did not neatly apply to paradigm moral properties, i.e. goodness and badness, rightness and wrongness. Instead, my argument for the existence of non-natural moral properties was more oblique. I started by identifying two types of non-natural normative properties: the properties of *being good for/bad for* some organism and role properties. I then noted that our moral judgments seem to be reliably influenced by these properties. Consequently, the following hypothesis is well confirmed: the function of the mechanisms that ground our dispositions to make moral judgments is to recognize non-natural normative properties. I argued that, if this hypothesis is true, we have good reason to believe that there are non-natural moral properties. While I would have preferred a more direct defense of ethical non-naturalism, the oblique approach I was forced to take is not without its perks. In particular, the defense of ethical non-naturalism I have offered nicely positions us to tie the capability of making veridical moral judgments to adaptive behavior. The evolutionary account I will offer does not take the form of a pre-established harmony account. Instead, I will argue that it would be straightforwardly selectively advantageous to recognize two distinct types of non-natural normative properties.

Consider role properties. Would it be evolutionarily advantageous to track these? The answer seems to be "yes." That Sam is a good parent, a good teacher, or a good hunter all seem to be clearly relevant to various decisions one might make with regard to Sam. If Sam is not a good parent, this gives one good reason not to want to procreate with Sam. If Sam is not a good teacher, this gives one good reason not to trust Sam with one's children. If Sam is not a good hunter, this gives one good reason not to entrust Sam with the task of procuring protein for one's tribe. Each of these decisions appears to be directly relevant to the probability that one will successfully pass on one's genes. I have argued that we have good reason to believe that role properties are non-natural. It further appears that recognizing role properties would be evolutionarily advantageous.

The case is even clearer with regard to the properties of *being good for so-and-so* and *being bad for so-and-so*. Knowing that *eating such-and-such a mushroom is generally bad for human beings* is clearly relevant to one's fitness. The same goes for knowing that *eating such-and-such an herb is generally good for humans*. Perhaps less obviously, it is also valuable to know what is good for and bad for other kinds of organisms. Thus, if one wants to establish successful agricultural practices, it will be important to know what *is good for* and what *is bad for*, e.g., one's corn and one's goats. There is little doubt that being able to track the *good for/bad for* properties for certain kinds of organisms would be evolutionarily advantageous.

If I am correct that there are straightforward evolutionary advantages to being able to track role properties and good for/bad for properties, the normative realist is in a position to fulfill both explanatory burdens. The two previous paragraphs explain the correlation between normative truth and adaptive success. Once this correlation has been explained, the realist is in a position to fulfill the original explanatory burden. What properties do our evaluative judgments track in virtue of which they are evolutionarily advantageous? Role properties and the good for/bad for properties.

Note, however, that this account has the same flaw as the pre-established harmony accounts offered by Enoch and Skarsaune. The account is enough to vindicate *normative realism* from the dual explanatory burden. It does not, however, offer any support for *moral realism*. A vindication of moral realism based on the evolutionary account I have offered must provide some reason to think that our moral judgments are, by and large, accurate.

In the previous chapter, I argued that role properties and good for/bad for properties are non-natural. I also argued that the function of our moral faculty is to recognize non-natural normative properties. In this chapter, I have argued that it would be evolutionarily advantageous to track role properties and good for/bad for properties. This puts us in a position to offer a very rough sketch of the evolution of our moral faculty. Simply put, our moral faculty *did not* evolve to track moral properties. Instead, our moral faculty originated as a way to track other non-natural normative properties, i.e. role properties and good for/bad for properties.

Consider our ability to do calculus. It is implausible to suppose that the ability to do calculus was, itself, the object of selection. More plausibly, a range of related cognitive abilities were evolutionarily advantageous. These cognitive abilities were selected for and, it is in virtue of possessing these cognitive abilities that we are capable of doing calculus. I propose a similar account of the evolution of moral cognition. On this account, our ability to track moral properties is a spandrel. That is, it was not itself selected for. Instead, our ability to recognize moral properties is a consequence of a more general ability: our ability to recognize non-natural normative properties. It was evolutionarily advantageous to have the ability to recognize moral properties even though we did not evolve this ability *in order to* recognize moral properties.

The evolutionary debunking argument: re-aimed

Setting the stage

I have offered a just-so story about the evolutionary history of our moral faculty. The just-so story, if true, offers to explain how it was evolutionarily advantageous to make evaluative judgments. Just-so stories are, however, of little interest. One needs some evidence that the etiological sketch in question is accurate. Unfortunately, it is no easy task to provide such evidence with regard to the account I sketched in the first half of this chapter. Central to the account is the claim that our evaluative judgments are formed in response to the presence of non-natural normative properties. It does not seem likely that one will find confirmation of this etiological account in, e.g., the fossil record. It is difficult to even imagine what sort of observational evidence might straightforwardly confirm the etiological hypothesis.

This does not, however, mean that there is nothing to be said in defense of the veracity of the evolutionary just-so story I have offered. Street makes the following observation about realist friendly etiological accounts like the one I have offered:

> The… thing to notice about this account is that it puts itself forward as a scientific explanation. It offers a specific hypothesis as to how the course of natural selection proceeded and what explains the widespread presence of some evaluative judgments rather than others in the human population… In putting itself forward as a scientific explanation, the… account renders itself subject to all the usual standards of scientific evaluation, putting itself in direct competition with all other scientific hypotheses as to why human beings tend to make some evaluative judgments rather than others. (Street 126)

On this point, Street is certainly correct. The just-so account I just sketched is subject to the usual standards of scientific evaluation. Many of these are straightforwardly observational. Unfortunately, for the reasons given above, I do not think much, observationally, can be said about the plausibility of the just-so story. Luckily, observational evidence is not the only way a theory can be confirmed. In the third chapter of the dissertation I sketched the general commitments of the "scientific worldview." A significant portion of that discussion was taken up by consideration of *inference to the best explanation* and the *super-empirical virtues*.

The picture went as follows: some theory, T_1, (in addition to auxiliary hypotheses) provides an explanation, E_1, of some set of observations, O. E_1 then competes with other explanations, $\{E_2, E_3, \ldots E_n\}$, where each explanation is provided by some theory (conjoined with a set of auxiliary hypotheses) in competition with T_1, $\{T_2, T_3, \ldots T_n\}$. If E_1 is the best explanation of O then we have some reason to think that T_1 is the correct theory. Some explanation, E_1, is better than another explanation, E_2, just in case E_1 has more super-empirical virtues than E_2.

The just-so evolutionary account I offered provides an etiological theory of the development of our moral faculty. It offers an explanation of the selective advantage conferred by our moral faculty. If this explanation is better than the explanations provided by competing theories, we have some reason to believe that the etiological account I have offered is correct. The remainder of this chapter will be dedicated to the project of giving reason to believe that the just-so story I have offered grounds the best explanation of the selective advantage conferred by our moral faculty. There are two parts to this project. First, I will argue that the anti-realist's explanation is not more virtuously parsimonious than the non-naturalist's explanation. Second, I will argue that the non-naturalist's explanation is more consilient than the explanation offered by the anti-realist.

Simplicity re-considered

One might think that the moral anti-realist is in a position to offer a simpler explanation of the evolution of our moral faculty than the just-so story I have so far sketched. The argumentative strategy is relatively straightforward. My just-so story holds that the moral faculty evolved to track non-natural normative properties. Consequently, I am committed to the existence of non-natural normative properties. The anti-realist can offer an etiological account that does not commit her to the existence of non-natural normative properties. This anti-realist explanation of the evolution of our moral faculty will be ontological simpler—I am committed to the existence of non-natural normative properties whereas the anti-realist is not.

I do not, however, think that this argumentative strategy on behalf of the anti-realist is likely to be successful. I have thus far argued that, in light of considerations regarding IBE and the cosmological limits on computation, role properties and good for/bad for properties are both mind-independent and non-natural. Even the moral anti-realist is committed to the existence of non-natural normative properties.

Nonetheless, the anti-realist can still claim that her evolutionary explanation is more parsimonious in that, though both the anti-realist and the realist must posit the existence of non-natural normative properties, at the very least, the anti-realist can do without positing the existence of non-natural *moral* properties.

In this sense, the anti-realist's explanation is more parsimonious than the explanation the realist has to offer. It is, however, not clear that the kind of simplicity on offer is of much value. Theories can be more or less simple in a number of ways.

Relevant here is a distinction between being more parsimonious in virtue of positing *fewer entities* and being more parsimonious in virtue of positing *fewer kinds of entities.* Intuitively, the latter type of simplicity counts as a virtue of a theory. There seems to be something illicit about unnecessarily cluttering our ontology with a range of metaphysically distinct *kinds* of entities. There does not, however, seem to be anything particularly virtuous about positing *fewer entities*. Consider a theory that posits the existence of n red objects as compared to a theory that posits the existence of n^2 red objects. There does not seem to be anything, in principle, less plausible about a theory that posits the existence of n^2 red objects than a theory that posits the existence of n red objects. Parsimony considerations push in the direction of preferring theories that posit fewer *kinds* of entities, but not in the direction of preferring theories that posit *fewer* entities.

It is not clear that simplicity considerations will be of much use for the anti-realist or the reductive realist. The claim was that, compared to the non-naturalist, the anti-realist (or reductive realist) could offer a simpler, and thereby superior, explanation of the etiology of our moral faculty. The plausibility of this claim will depend on how fine-grained one prefers one's distinctions between types of properties. If the arguments I offered in chapter four and five are effective, everyone is committed to the existence of non-natural normative properties. If one takes the set of non-natural *moral* properties to be merely a subset of the set of non-natural normative properties, then the anti-realist's explanation is simpler in that it posits *fewer entities of the same kind.* But this kind of simplicity is not a theoretical virtue; it is not a guide to which explanation is best.

The moral anti-realist and the reductive realist are committing to thinking that they will eventually be able to offer an account of the etiology of our moral faculty that does not make reference to non-natural moral properties. Even if such an account is forthcoming, it is not obvious that it will be a better explanation than the one the non-naturalist hopes to offer. The anti-realist's explanation does not instantiate the kind of simplicity that counts as evidence for the superiority of an explanation.

The evolution of disgust

The above considerations are largely defensive. A more interesting question is: what kind of positive reasons do we have for accepting the explanatory sketch I offered at the outset of this chapter? In what remains of this chapter I will argue that the best explanation of certain aspects of our moral faculty is that the moral faculty (in part) evolved to recognize the good for/bad for properties. If successful, I will have simultaneously achieved two distinct aims. First, I will have provided reason to believe that the moral faculty evolved to recognize non-natural normative properties other than the right and the good. Acceptance of this claim was central to my defense of normative realism from Street's evolutionary argument. If I can show that the best explanation of certain aspects of our moral faculty is that the moral faculty (in part) evolved to be responsive to instantiations of the good for/bad for properties, I will have thereby provided some reason to believe that the just-so story I offered in response to Street's argument constitutes an accurate account of our evolutionary history. Second, if our moral faculty evolved to recognize instantiations of the good for/bad for properties, we have good reason to believe that the function of whatever grounds our dispositions to form moral judgments is to recognize non-natural normative properties. Defense of this thesis about the function of our moral faculty was essential to the success of the argument for ethical non-naturalism I developed in the fifth chapter.

As has often been the case in this dissertation, I will start somewhat far afield from my eventual goal. Instead of immediately considering the moral faculty, I will start by considering a *prima facie* tangentially related phenomenon: poison disgust. I will argue that the best explanation of certain aspects of our disgust response is that portions of our disgust response evolved to track the good for/bad for properties. I will then point to evidence that suggests that poison disgust is part of our moral faculty. Finally, I will argue that the best explanation of certain phenomena is that we have a single faculty responsible for recognizing diverse non-natural normative properties. The first step is to specify which aspect of our disgust response I am interested in explaining.

It is widely accepted that the set of affective responses that fall under the umbrella of the moniker "disgust" are caused by (at least) three distinct mechanisms (Olatunji et al. 1243).[39] A surprising variety of stimuli appear to be capable of causing disgust. Consider the following three, largely unrelated, phenomena: the taste of an extremely bitter drink, the thought of a brother and sister sleeping together, and the smell of rotting carrion. Each causes disgust; however, the three causes of disgust appear to be importantly different. This variety in disgust stimuli suggests two distinct explanatory strategies we might take. We might try and offer a *unifying* explanation of the cognition that underlies disgust. That is, we could hold that there is something importantly similar about the taste of an extremely bitter drink, the thought of a brother and sister sleeping together, and the smell of rotting carrion such that our disgust response towards each is the consequence of a single cognitive mechanism. Alternatively, we could offer a *divorcing* explanation of the cognition that underlies disgust. That is, we could hold that three (or two) distinct mechanisms are responsible for our disgust response. Each mechanism responds to a different property; our disgust response to these disparate phenomena is not best explained by some hidden unity amongst the stimuli.

[39] For those trained in the traditional methods of philosophy, the following discussion may seem imprecise. What exactly is meant by "disgust?" One might wish for a conceptual analysis or, minimally, a careful phenomenological description. The literature I am engaging with here offers neither. The referent(s) of "disgust" are most easily identified via ostension: the mental state(s) one tends to have when one displaying the gape face. (See page 199 and 200 for a brief description of the gape face.)

Recent work on disgust suggests that we can helpfully distinguish three subsystems that, conjointly, constitute the systems responsible for our disgust response. That is, recent research suggests that we ought to accept a divorcing explanation of the cognition that underlies disgust (Olatunji et al.). I will distinguish between three types of disgust then differentiate between the relevant cognitive mechanisms by noting the kind of disgust each mechanism appears to be responsible for.

Label the first of the three types of disgust "*moral disgust*." Moral disgust tends to lead us to make a moral judgment. Thus, when we feel disgust at the thought of a brother and sister sleeping together, the disgust tends to co-vary with a judgment that *they (morally) ought not do that!* Other examples of moral disgust are the feeling of disgust some people get when they vividly imagine homosexual intercourse, or the feeling of disgust one might get when imagining an act of necrophilia. Perhaps surprisingly, I will dedicate very little time to considering moral disgust. My aim is to show that the best explanation of the evolution of our moral faculty involves reference to non-natural normative properties. This only requires that I demonstrate that *one* kind of disgust is part of our moral faculty and evolved to track non-natural normative properties. I set aside moral disgust because I do not need it to make my case.

The same can be said for the second kind of disgust: *parasite disgust*. Parasite disgust "evolved to provide one way to protect against infection from pathogens and parasites, namely, by avoiding them" (Kelly 48). Parasite disgust is the type of disgust we feel when we smell rotting carrion, consider bathing in feces, or consider touching a horribly pockmarked conspecific. Again, I have very little to say about the parasite avoidance mechanism. I will set it aside.

I am primarily interested in the final kind of disgust and the cognitive mechanism that underlies it. Call this kind of disgust, "*poison disgust*" and the mechanism that underlies it the "*poison mechanism*" (Kelly 48). Poison disgust is the kind of disgust caused by noxious tastes, e.g. the taste of an extremely bitter drink.

Rozin's pioneering work on disgust suggests that the poison mechanism evolved as a consequence of the "omnivore's dilemma" (Rozin). The problem is relatively straightforward: "[An] organism must eat, but it must be selective in what it consumes, because many things that seem edible are actually harmful when ingested…" (Kelly 46). The problem is particularly relevant for omnivores. While the carnivore's digestive system can specialize in immunity from meat-related poisons and the herbivore's digestive system can specialize in immunity from plant-related poisons, the omnivore's digestive system must remain unspecialized. In lieu of evolving a digestive mechanism capable of handling a wide range of toxins, we can expect omnivores to have evolved a method of poison avoidance:

> One common way to navigate the problems raised by the omnivore's dilemma is provided by acquired taste aversions. These provide a way to narrow down culinary options by implementing a "once bitten, twice shy" rule. In humans, this variety of "shyness" manifests as an intense aversion directed toward the offending food type. Thus, a type of food that has induced sickness in the past comes to be avoided in the future. (Kelly 46)

An important evolutionary function of disgust is the avoidance of the ingestion of harmful material. The poison mechanism underlies this function of disgust (Kelly).

Earlier I argued that any account of the evolution of our moral faculty will have to identify the set of external world properties our moral judgments are primarily responsive to. The argument for this conclusion was quite simple: if our moral faculty was the product of selection, it must be the case that the moral faculty improved fitness. However, if the moral faculty is to improve fitness, it must get us to respond appropriately to the external world because selective pressures are only to be found in the external world. Consequently, if the moral faculty was the product of selection, it must be the case that the moral faculty is responsible to some set of external world properties.

The schema applies equally well to the poison mechanism. If the poison mechanism was conducive to fitness, it must be the case that the poison mechanism was responsive to some set of external world properties. Any explanation of the evolution of the poison mechanism must identify some set of external world properties the mechanism is responsive to.

The reader likely already knows where my argument is headed. I intend to argue that the poison mechanism evolved to track a non-natural normative property, in particular, the property of *tending to be bad for humans*. This explanatory strategy contrasts with one whereby one holds that the poison mechanism evolved to track some set of non-normative properties.

Fessler and Machery have, recently, presented a novel account of an evolutionary relationship between culture and cognition. They start by noting that a wide variety of adaptive behavior is learned, not innate. This observation gives rise to the following explanatory challenge:

> Two principal obstacles confronting learners who seek to benefit from others' knowledge are the richness of the informational environment and the incompleteness of the discernable information therein. First, human behavior is enormously complex, varying across contexts and persons, while linguistic utterances convey information ranging from the trivial to the life-saving. If, as is often tacitly presumed, learners were indiscriminate sponges, then (1) learners would often fail to understand how to apply what they learned, and (2) learners would fail to properly prioritize their acquisition efforts, often resulting in both precocity in domains irrelevant to the learner and retardation in relevant domains. Second, much social learning involves the problem of the poverty of the stimulus, and many actions and utterances explicitly present only fractional portions of the information that motivates them. (Fessler and Machery 513)

The problem Fessler and Machery consider is closely related to Chomsky's seminal criticism of behaviorism (Chomsky). Unsurprisingly, the structure of the solution Fessler and Machery propose is closely related to the explanatory move made by Chomsky: "We suggest that, for many domains of learning, natural selection has addressed both of these problems by endowing the mind with [domain specific] inborn mechanisms, possessing considerable content, that serve to structure the acquisition of cultural information" (Fessler and Machery 513).

Fessler and Machery next consider the kind of conditions under which one could expect evolution to endow organisms with inborn mechanisms for the acquisition of cultural information. Fessler and Machery identify three conditions that must hold for a domain-specific cultural information acquisition mechanism to have evolved:

> First, the domain must have been of substantial and relatively uniform importance to biological fitness across the diverse socioecological circumstances that characterized ancestral human populations, as this will have provided the steady selection pressures necessary for the evolution of a complex adaptation. Second, the domain needs to involve content that will have varied significantly across said circumstances as, on the one hand, this precludes the evolution of extensive innate knowledge, and, on the other, this maximally exploits the culture's ability to effectively compile information of parochial relevance. Lastly, the domain must be one in which individual learning through trial-and-error or direct observation would have been either very costly or impossible much of the time under ancestral circumstances. (Fessler and Machery 514)

There may be selective pressures that are relatively steady across environments but manifest themselves in substantively different ways across each environment. The fact that these selective pressures are independent of any given environment gives us some reason to think that organisms will evolve a mechanism for dealing with these selective pressures. However, the fact that these selective pressures manifest themselves in different ways depending on the specific environment in which they occur gives us some reason to believe that any evolutionary solution will have to be flexible enough to apply to a broad range of different manifestations of the same selective pressures. There is an obvious candidate for an evolutionary solution: learning. The general ability to modify one's behavior to suit the environment will allow an organism to respond to selective pressures that are invariant across environments yet manifest themselves differently per environment. However, there will be a subset of such selective pressures such that a general ability to modify one's behavior to suit the environment will not provide an adequate solution. Discovering a behavioral solution to some selective pressures can be extremely costly. There are at least two ways in which learning new behavior can be costly. First, the method of trial and error may be too likely to lead to death (or sterilization). In those instances where getting it wrong can be fatal, it is unlikely that the general ability to modify one's behavior to suit the environment will provide an adequate response to the selective pressures in question. Second, information acquisition may simply take too long. Discovering a behavioral solution to the selective pressures in question may take the majority of an organism's lifespan. If this is the case, organisms will remain entirely vulnerable to the selective pressures in question through the majority of their lifespan. Some other solution is needed is reponse to these selective pressures.

Fessler and Machery have already suggested what such a solution might be: the ability to learn from the behavior of those around them. For the reasons discussed above, we can expect this ability to require the development of cognitive mechanisms with significant innate content.

This discussion in relatively abstract; it is worth considering an example of the kind of cognitive mechanism Fessler and Machery have in mind. They offer the following discussion:

> Skill in using tools is a determinant of fitness in all known small-scale societies. Some of the informational basis of tool use skill can be obtained through trial-and-error learning, as artifacts' affordances will often bias experimental efforts in the direction of techniques congruent with the tool's design. However, in many instances, trial-and-error learning will be more expensive (in terms of time, and in terms of risk of injury to self, the tool, or other objects or persons) than social learning; even in technologically simple societies, acquiring mastery of some tool techniques (e.g., flintknapping) is laborious, or even impossible, without cultural information. We can therefore expect natural selection to have crafted domain-specific cultural information acquisition mechanisms dedicated to this domain… These mechanisms may contain conceptual primitives, such as "piercing tool," "cutting tool," "lever," "carrying tool," "container," and so

> on, that aid in the acquisition of cultural information linking a specific tool, a specific objective, and a specific technique. (Fessler and Machery 516)

Are there other domains in which we can expect humans to have evolved a domain-specific cultural information acquisition mechanism? Apropos to our project in this chapter, it appears that we can expect humans to have evolved a domain-specific cultural information acquisition mechanism regarding the ingestion of food. There are three criteria a selective pressure must meet in order to be a candidate for the evolutionary cause of a domain-specific cultural information acquisition mechanism. First, the selective pressure must be prevalent across environments. Second, the selective pressure must manifest itself in different ways across environments. Finally, the cost of trial-and-error learning must be high.

The selective pressure constituted by the need to avoid eating harmful foods fulfills all three criteria. Independent of environment, organisms, and omnivores in particular, need a method for avoiding the ingestion of harmful substances. There will be significant variance, across environments, in the markers of a substance that is harmful to ingest. Finally, the cost of trial-and-error learning regarding the ingestion of potentially harmful substances is extremely high. Ingestion of the wrong substance can easily be fatal. It appears that we have some reason to believe that humans would have evolved a domain-specific cultural information acquisition mechanism to facilitate the task of avoiding the ingestion of harmful substances.

There is further reason to believe that poison disgust is, at least in part, the consequence of some such mechanism. If poison disgust is, at least in part, the consequence of a domain-specific cultural information acquisition mechanism, we would expect to find the disgust response playing a role in the transmission of cultural information.

It is widely accepted that disgust is associated with a specific facial expression—the “gape” face:

> Darwin's … initial description of the facial expression of disgust emphasized the gape (i.e., mouth held open widely) but also mentioned a raised upper lip and a nose wrinkle. Izard's (1971) description of the portrayal of disgust included a raised upper lip, the mouth corners drawn and back, and the tongue forward and may be slightly protruding… Ekman and Friesen's (1975) description of the disgust expression consists of a lip retraction, a raised lower lip, and a wrinkled nose. This facial expression often signals the experience of nausea, increased salivation and parasympathetic responding and functions to protect the body from the ingestion of an object… (Elwood and Olatunji 101)

Disgust comes equipped with a unique signaling system. Importantly, the gape face appears to be largely involuntary; people communicate disgust even if they do not intend to and even if they are unaware that they are gaping:

> [P]eople who are only very slightly disgusted tend to gape even if the facial movements are so minimal and understated that the gaper is not aware of producing them. These *microexpressions* flash quickly across the face; they last as briefly as forty milliseconds and can remain completely below the conscious awareness of the individual producing them… People also reveal their disgust even when they would rather keep it to themselves. When genuinely disgusted, people tend to flash a microgape even [when] they try to voluntarily suppress it. As nice as it would be to be able to hide it sometimes—for instance, while choking down a nasty dinner at the in-laws'—this kind of "leakage" betrays disgustedness by displaying it on the face, subtly but automatically… (Kelly 65)

When disgusted, one will signal disgust, whether or not one intends to. If the poison mechanism is part of a domain-specific cultural information acquisition mechanism, we would expect the poison mechanism to be associated with some method of communication. Essential to communication is the transmission of a signal. It appears that poison disgust, by way of the gape face, reliably co-varies with the transmission of a signal.

Transmitting a signal is not, of course, enough for successful communication. It must also be the case that the signal is received. Not only is the gape face universally recognized, the perception of the gape face causes disgust in the perceiver:

> Not only are people able to naturally recognize a gape *as* an expression of disgust, but doing so often involves the extra step of actually becoming disgusted oneself. This is striking. Not

> only is recognition of disgust automatic, but the processes involved automatically put the recognizer into a similar mental state as the person being observed. (Kelly 66)

It appears that there is cross-cultural recognition of the gape face as an indicator of disgust. Furthermore, perceiving the gape face can cause the perceiver to feel disgust.

If disgust is part of a mechanism that allows humans to acquire cultural information regarding what not to eat, it must be the case that we have some way of knowing when conspecifics are disgusted. It appears that the gape face reliably serves to communicate that someone is disgusted. Mere communication of information regarding when a conspecific is disgusted is, of course, not particularly good evidence that disgust is part of a domain-specific cultural information acquisition mechanism. If disgust is part of a domain-specific cultural information acquisition mechanism it must be the case that communication regarding the disgust of conspecifics can influence what humans are willing to eat. It appears that the disgust response to certain stimuli is often learned from conspecifics:

> [F]rom the population-level point of view, the distribution of disgust elicitors realizes a familiar pattern of within-group similarity and between-group differences… This pattern suggests that there is an important role for social learning and acquisition in what disgusts individual people. In other words, disgust elicitors are not all innately specified, nor are they all acquired via individual learning. (Kelly 93)

Culture appears to be an important repository for information regarding the appropriate objects of disgust.

In light of the contours of the omnivores dilemma, we have reason to expect humans to have evolved a domain-specific cultural information acquisition mechanism to aid in the avoidance of the ingestion of harmful substances. The poison mechanism appears to be a prime candidate for such a domain-specific cultural information acquisition mechanism. Not only does it appear to play the appropriate role in modifying behavior, disgust behavior appears to be transmitted culturally. There is at least some reason to believe that the poison mechanism is part of a domain-specific cultural information acquisition mechanism. In the remainder of this chapter I will assume that this is the case. I do not take myself to have presented definitive, or even compelling, evidence that the poison mechanism is part of a domain-specific cultural information acquisition mechanism. There is some reason to believe the hypothesis; however, the empirical verdict is still out. The structure of my argument in the remainder of this chapter should be familiar: I will assume that the preponderance of *a posteriori* evidence points in a specific direction and then see, philosophically, what follows.

The preceding discussion was set up by the following observation: if the poison mechanism was conducive to fitness, it must be the case that the poison mechanism was responsive to some set of external world properties. Any explanation of the evolution of the poison mechanism must identify some set of external world properties the mechanism is responsive to. Earlier in this chapter I offered a just-so story whereby the moral faculty evolved to track non-natural normative properties. In defense of this just-so story, I will argue that poison disgust is part of the moral faculty and that the poison mechanism evolved to be responsive to the non-natural normative property of *tending to be bad for humans*. This hypothesis about the evolutionary function of disgust contrasts with those hypotheses that hold that the poison mechanism evolved to track some set of non-normative properties, e.g. extreme bitterness or extreme sourness. How does my etiological hypothesis fare?

Assume that the poison mechanism evolved as part of a domain-specific cultural information acquisition mechanism. Consider, again, the first two conditions that must be in place for a domain-specific cultural information acquisition mechanism to evolve: (1) there must be a selective pressure prevalent across environments and (2) that selective pressure must manifest itself in different ways across environments. What selective pressure might be responsible for the evolution of the poison mechanism? It appears to be the case that the poison mechanism evolved to prevent humans from ingesting harmful substances. But notice that "harmful" is just another way of saying "bad for humans." So, put another way, the poison mechanism evolved to prevent humans from eating things that were *bad for them*. Note how nicely the property of *being bad for humans* fulfills the above two necessary conditions for the evolution of a domain-specific cultural information acquisition mechanism. In every environment there will be things that are bad for humans to eat. The specifics of what is bad for humans to eat will vary per environment.

Consider the following hypothesis: poison disgust evolved to prevent humans from eating things that were bad for them. The hypothesis appears to fare nicely. It identifies a property that would be fitness conducive to be responsive to and the property can help explain why poison disgust evolved as a domain-specific cultural information acquisition system.

How do competing hypotheses fare? To the best of my knowledge, no one has attempted to offer an evolutionary explanation of the poison mechanism that does not rely on the notion of substances that are *harmful* for humans to digest. This makes it difficult to assess competing hypotheses; however, there are in principle reasons to think that any evolutionary hypothesis that fails to rely on the non-natural normative property of *being bad for humans* will fare poorly. The crux of the problem is that the poison mechanism evolved as part of a domain-specific cultural information acquisition mechanism. Thus, any candidate for an etiological explanation of the poison mechanism must identify a set of properties that the poison mechanism is responsive to such that (1) being responsive to this set of properties explains why poison disgust was fitness conducive, (2) the set of properties is invariant across environments, and (3) the set of properties manifests itself differently across environments.

There are, broadly speaking, two strategies one might try and take in identifying a property that fulfills the above three conditions. One might attempt to identify a single property that can do all of the necessary explanatory work *or* one might attempt to identify some disjunctive set of properties that can do all of the necessary explanatory work. Note that, if one takes the former approach, an explanation that relies on the property of *being good for humans* will be comparatively simpler than competing explanations. We already have reason to believe that the property of *being good for humans* constitutes an ontologically robust part of the fabric of our universe. Relying on the property to explain the evolution of poison disgust comes at zero ontological cost.

It is difficult to imagine some other single property that could do all of the needed explanatory work. I suspect that any other term someone uses to identify a property that is putatively capable of doing the necessary explanatory work will just be another way of talking about the property of *being good for humans*. But suppose that this is not the case. Whatever putative property one cooks up, if it is going to explain the evolution of the poison mechanism, we will have to add it to our ontology. Explanations of the evolution of the poison mechanism that rely on the single property of *being bad for humans* are likely to be more ontologically parsimonious than competing explanations that attempt to explain the same phenomena in terms of some other single property.

But consider the latter strategy. One might attempt to explain the evolution of the poison mechanism in terms of some disjunction of properties. It is unlikely that this strategy will be effective. One might take two distinct approaches in attempting to make good on the strategy: one might attempt to identify a disjunction of (conjunctions of) microphysical properties or one might attempt to identify a disjunction of properties at some higher level.

Consider the first strategy. Conditions (2) and (3) on a successful explanation of the evolution of the poison mechanism rule out reliance on any property composed of disjuncts of (conjuncts of) microphysical properties; in any sense in which a disjunctive property fulfills (2), it fails to fulfill (3). Alternatively, in any sense in which a disjunctive property fulfills (3), it fails to fulfill (2). The point is perhaps best made via illustration. Consider some toy disjunction of conjunctions of microphysical properties, where each M_x picks out a conjunction of microphysical properties: $\{M_1 \vee M_2 \vee \ldots \vee M_n\}$. Call this disjunctive property MP. If MP is going to help explain the evolution of the poison mechanism, MP must be invariant across environments and MP must manifest itself differently across environments.

MP manifests itself differently across environments; only some members of the set of MP will be present in any given environment. In virtue of some member of the set of MP being in every environment in which humans might have evolved, MP is also invariant across environments. However, note that the invariance of MP across environments *cannot* explain the evolution of the poison mechanism as a domain-specific cultural information acquisition mechanism. The problem is that, though MP is invariant across environments, MP does not ground an invariant *selective-pressure* across environments. Consider some environment, E_1. Suppose some subset of MP is present in E_1: $\{M_1, M_2, \ldots M_{10}\}$. In E_1, the selective pressure is to avoid M_1 through M_{10}. In some other environment, E_2, the selective pressure may be to avoid M_{11} through M_{23}. Thus, MP fails to be invariant across environments in a way that would explain the evolution of a domain-specific cultural information acquisition mechanism.

One might, however, opt for an explanation that relies on a disjunctive property that does not consist of conjunctions of microphysical properties. Consider the following toy example: *being the cause of gastrointestinal distress* ***or*** *being the cause of blindness.*[40] Call the above disjunctive property GB. As per the arguments in the fourth chapter, the two properties that constitute GB are likely non-natural: GB is a disjunctive property but is not composed of (conjuncts of) microphysical properties. Note that GB is both invariant across environments and will manifest itself differently in different environments. Thus, it seems that GB can plausibly ground an explanation of the

[40] We likely need to specify the properties a bit more, so that they only apply to the ingestion of substances; however, the strategy can be illustrated well enough without such specification.

evolution of the poison mechanism. What can be said in favor of my preferred explanation when compared with an explanation that relies on GB?

In brief, the two explanations do not appear to be in competition. If the ingestion of some substance causes either gastrointestinal distress or blindness then it is likely that, under most circumstances, ingestion of the substance is *bad for humans*. Given the sort of substances it appears poison disgust evolved to help us avoid, no matter how one attempts to cash out the properties that the poison mechanism evolved to help us avoid, it appears that these properties will always instantiate the property of *tending to be bad for humans*. Thus, any plausible explanation of the evolution of the poison mechanism will result in the poison mechanism having evolved to help humans respond to a non-natural normative property.

It is important to see why this move on the part of the non-naturalist is not a copout. Properties such as *being the cause of gastrointestinal distress* are multiply realizable. Consequently such properties can be invariant across environments. It is plausible to suppose that one can account for the evolution of the poison mechanism by offering an enormous (presumably infinite) disjunction of multiply realizable properties. I have argued that the non-naturalist can be happy with this disjunctive style of explanation because any such explanation will merely be a rephrasing of an explanation in terms of the properties *being good for humans/being bad for humans*. From the perspective of traditional non-naturalist, this is a very surprising claim. I have embraced the possibility of providing an (disjunctive) analysis of a non-natural property—the traditional non-naturalist is committed to denying the existence of any such analysis. My version of non-naturalist has no such commitments. The properties of *being good for/being bad for* can be non-natural even if they are analyzable. I am only committed to denying that non-natural properties are *token-token identical* with microphysical properties. Consequently, I can happily accept any attempt to reduce normative properties so long as they are not (token-token) reduced to microphysical properties.

We now have some reason to believe that the poison mechanism evolved to help humans respond to a non-natural normative property. This provides an important piece of the evolutionary puzzle; however, I am primarily interested in providing a defense of a different evolutionary hypothesis: the *moral faculty* evolved to recognize non-natural normative properties. More needs to be said to link this evolutionary claim about the poison mechanism to an evolutionary claim about the moral faculty.

Disgust and the evolution of the moral faculty

My aim in the second half of this chapter is to give the reader some reason to believe that [H] is true: the function of the mechanisms that underlie our dispositions to make moral judgments is to recognize non-natural normative properties. I have so far argued that poison disgust evolved as a way to recognize the property of *being bad for humans*. If poison disgust is part of our moral faculty, we have good reason to believe [H].

A growing literature documents the tight connection between disgust and moral judgments. It is now widely accepted that one can increase the severity of subjects' moral judgments by inducing disgust. In a study that has become a social psychology classic, Wheatley and Haidt used post-hypnotic suggestion to cause subjects to experience "a brief pang of disgust… a sickening feeling in … [their] stomach" when they read a particular word (Wheatley and Haidt 780). Some subjects were conditioned to feel disgust when they read the word "often," others were conditioned to feel disgust when they read the word "take." Subjects then read vignettes involving moral transgressions. Some vignettes included the disgust cue word; others did not. Wheatley and Haidt summarize their findings:

> Participants found moral transgressions to be more disgusting when their hypnotic disgust word was embedded within the vignettes than when this word was absent. Moreover, the disgust word caused participants to rate transgressions as more morally wrong. Apparently, participants used their feelings of disgust (attached only to a word, not to the act in question) as

> information about the wrongness of the act. (Wheatley and Haidt 781)

One might be worried that these findings suggest a relationship between moral disgust and moral judgments, or parasite disgust and moral judgments, but not between poison disgust and moral judgments. We need evidence that more straightforwardly relates poison disgust with moral judgments.

It has been shown that inducing poison disgust (via ingestion of a bitter liquid) leads subjects to judge moral violations more harshly than control subjects. Subjects who drank a bitter liquid before ranking the severity of moral violations were likely to judge the moral violations more severe than were subjects who drank a neutrally flavored beverage (Eskine, Kacinik and Prinz e41159). Even more surprising is that the relationship appears to be bi-directional: "reading about moral transgressions… resulted in inducing gustatory disgust…" (Eskine, Kacinik, Webster e41159). Subjects who considered moral transgressions just prior to drinking a neutrally flavored beverage were more likely to find the beverage disgusting than were control subjects.

Some etiological account of this bi-directional relationship is owed. Why would there be this kind of tight connection between moral judgments and disgust? Were an explanation of the etiology of our moral faculty capable of offering insight into an otherwise surprising phenomenon, it would be a significant mark in its favor. Can the etiological sketch I offered at the outset of this chapter offer an explanatory grip on this bi-directional relationships? The answer is "yes."

I have so far argued that our moral faculty evolved, not to track moral properties, but to track the non-natural normative properties *being bad for humans*. Instead of having a distinct set of competencies for responding to each type of non-natural normative property, we have a single faculty responsive to all non-natural normative properties. If this account is correct, we should expect the output of our moral faculty to be influenced, not only by moral properties, but also by a variety of non-natural normative properties.

The bi-directional relationship between disgust and moral judgments can be explained by a third factor: the property of *being bad for humans*. Poison disgust evolved to help humans avoid eating things that were bad for them. If the just-so etiological account I sketched in the first half of this chapter is correct, it should come as little surprise that feeling disgust would influence one's moral judgments. Our moral faculty was not designed to flawlessly track moral properties. Instead, it tracks a cluster of properties. We would expect the presence of any of the appropriate properties to influence the output of the faculty. Poison disgust is a signal that a negatively valenced non-natural normative property is present. On the etiological sketch of the moral faculty I offered, the presence of any non-natural normative property should influence our moral judgments.

Perhaps more interesting is the etiological hypothesis' ability to explain why moral judgments can influence how things taste. From a purely physiological perspective, this phenomenon is extremely surprising. It is tempting to think that taste is nothing more than a response to the chemical make-up of whatever it is that one is tasting. It appears, however, that this account will not do. Something needs to be said regarding why moral judgments can influence how food tastes.

The view that the poison mechanism is part of a domain-specific cultural information acquisition mechanism can do significant explanatory work here. It is important that our poison disgust response be modifiable; if it is not modifiable, then one cannot learn about the appropriate targets of disgust from one's conspecifics. We would expect that judgments about how healthy something is would have (at least some) influence on how it tastes. (Though obviously not very much influence; we've all had to choke down our share of medicine.) Now suppose that the just-so story I sketched at the outset of this chapter is accurate. We do not have a distinct capacity to respond to each kind of non-natural normative property. We are now nicely positioned to explain why making moral judgments can change how something tastes. We would expect judgments about the non-natural normative property of *being bad for humans* to have an influence on how something tastes; however, our capacity for responding to non-natural normative properties is not particularly fine-grained. Just as the presence of the property of *being bad for humans* can bleed over into our moral judgments, the presence of moral properties can bleed over into the taste of our food.

The ethical non-naturalist's etiological explanation of the nature of our moral faculty has a surprising upshot: it offers an explanation of entirely disparate phenomena. This is a valuable result for the non-naturalist. Consilience is an important super-empirical virtue. In virtue of being able to relate surprising observations made by cognitive scientists to an evolutionary account of the etiology of our moral faculty, the non-naturalist's explanation of the evolution of our moral faculty appears to be relatively consilient. Not only is the non-naturalist in a position to offer an evolutionary explanation of the nature of our moral faculty, the explanation is super-empirically virtuous.

Neither the anti-realist nor the reductive realist is in a position to offer a competing etiological explanation of our moral faculty. This makes it difficult to compare the competing explanations to determine which explanation is better. Nonetheless, there are some reasons to think that any explanation the anti-realist is able to provide will not be able to offer a unified account of *both* the evolution of morality *and* the bi-directional causal relationship between gustatory experiences and moral judgments. If this is the case, it constitutes a significant mark in favor of the non-naturalist's explanation.

Consider the structure of the unifying explanation I have offered. Two distinct phenomena need explanation: the evolution of our moral faculty and the surprising bi-directional relationship between gustatory displeasure and moral judgment. The explanation I have offered was unifying insofar as it identified an underlying factor that played an ineliminable role in explaining all of these phenomena: non-natural normative properties. A theory offers to provide unifying explanations insofar as it shows that a shared set of properties plays a fundamental role in the explanation of multiple phenomena.

It is difficult to see how the anti-realist or the reductive realist will be able to arrive at any similar unity of underlying properties. In order to provide a unifying explanation, my meta-ethical opponent will have to point to some external world properties that (1) play an ineliminable role in an explanation of why it was evolutionarily advantageous to make moral judgments and (2) play an ineliminable role in an explanation of the bi-directional relationship between gustatory displeasure and moral judgments. The problem is that both the anti-realist and the reductive realist think that our moral judgments are primarily responsive to nothing over-and-above some set of microphysical properties. Thus, the external world properties that will putatively play the unifying role must be some microphysical properties. But consider the striking difference between the sort of properties that tend to cause moral judgments, e.g. the suffering of kin, and the sort of properties that cause gustatory displeasure, e.g. bitterness. If one can point to nothing over-and-above the microphysical properties of bitterness and the suffering of kin, it is *extremely* difficult to imagine what property might be shared between the two that can serve in an explanation of both the etiology of our moral faculty *and* the bi-directional relationship between moral judgments and gustatory displeasure.

Nothing I have said here amounts to an argument with the conclusion that an anti-realist unifying account is impossible. I have merely pointed out the *prima facie* difficulty the anti-realist faces in any attempt to provide such an explanation. We ought not rule out the possibility of an anti-realist friendly explanation that can find a commonality between these apparently unrelated phenomena; however, were I a betting man, I'd bet against the anti-realist.

The non-naturalist can now turn the explanatory tables on the anti-realist and the reductive realist. Suppose that the anti-realist or reductive realist is capable of coming up with a competing explanation of the evolution of our moral faculty. It is likely that this explanation will not be capable of explaining *both* the evolution of our moral faculty *and* the bi-directional relationship between moral judgments and gustatory displeasure. The upshot: it appears that the non-naturalist's etiological explanation of the development of our moral faculty has an important super-empirical virtue lacked by any explanation that is likely to be offered by her meta-ethical competitors. This gives us some reason to believe that the non-naturalist's explanation is the best of all competing explanations, which in turn gives us some reason to believe that ethical non-naturalism is an accurate theory.

In the first half of this chapter I offered the following just-so story of the evolution of our moral faculty: our moral faculty did not evolve to track moral properties; instead, our moral faculty evolved to recognize non-natural normative facts. Our ability to recognize moral facts is a fortuitous accident. We now have some reason to believe that this just-so story is accurate. We have independent reason to believe that the property of *being bad for humans* is a non-natural normative property. It appears that the poison mechanism evolved as a way to get humans to avoid ingesting harmful substances. Put another way, it appears that the poison mechanism evolved as a way to keep track of a non-natural normative property—the property of *being bad for humans*. Consideration of the etiology of the poison mechanism gives us some reason to think that we evolved a general capacity to be responsive to non-natural normative facts.

If our moral faculty is a consequence of a more general ability to respond to non-natural normative facts, we would expect to see a tight connection between our moral judgments and our judgments about other non-natural normative properties. The evidence for a tight relationship between disgust and moral judgments is ever growing. In particular, it appears that the experience of poison disgust is importantly related to the severity of our moral judgments. This kind of entanglement between moral judgments and judgments about other kinds of non-natural normative properties is exactly what we would expect if our ability to be responsive to moral facts is a lucky byproduct of our ability to be responsive to non-natural normative facts more generally. The just-so story I offered in the first half of this chapter is beginning to look increasingly plausible.

Conclusion

I hope to have accomplished two distinct goals in this chapter. First, I hope to have sketched an etiological account of the evolution of our moral faculty. The success of this etiological account involves answering two distinct but related explanatory burdens. First, I have offered an explanation of the relationship between normative truths and selective pressure. Second, I have identified a set of external world properties that our moral judgments are responsive to, i.e. non-natural normative properties. In offering this just-so evolutionary story, I take myself to have provided a first pass at defending ethical non-naturalism from Street's criticism. In the latter half of the chapter I argued that there is reason to believe that the non-naturalist's explanation of the selective advantage conferred by our moral faculty is better than any explanation the non-naturalist's meta-ethical competitors will be able to offer.

The importance of this chapter of the dissertation ought not be understated. In the latter half of this chapter I argued that the non-naturalist's etiological explanation is a candidate for the best explanation because it offers to unify phenomena that previously appeared unrelated. If the non-naturalist's etiological explanation is better than the competitors, we have a reason to think that ethical non-naturalist is this case and this reason is independent of the arguments I offered in the fourth the fifth chapters. More important, however, is the relationship that holds between the defense of ethical non-naturalist I offered in the fourth and fifth chapters and the non-naturalist's etiological explanation developed in this chapter. The etiological account developed in this chapter draws heavily on a hypothesis (briefly) defended in the fifth chapter: [H] the function of the mechanisms that ground our dispositions to form moral judgments is to recognize non-natural normative properties. In the fifth chapter I argued that, in light of a range of explanatory considerations explored in the fourth and fifth chapters, [H] is highly confirmed. In this chapter, we have seen that [H] can ground an evolutionary explanation of the development of our moral faculty. This counts as a *significant* mark in favor of [H]. It was a mark in favor of the non-naturalist's etiological account that it could explain the bi-directional relationships between moral judgments and disgust. Analogously, it counts in favor of [H] that it can ground an evolutionary explanation of the development of our moral faculty. As an explanation of the role of our moral faculty, [H] is *strikingly* successful. Not only does it account for the phenomena discussed in chapters four and five, it grounds an evolutionary explanation of the development of our moral faculty and thereby also explains the two bi-directional relationships discussed above. The harder we look, the more we find that non-natural moral properties play a role in the explanation of a surprising and ever-growing range of phenomena.

CHAPTER SEVEN:
MORAL DISAGREEMENT

Introduction

In the previous chapter I considered an attempt by the anti-realist to demonstrate that moral realism and the scientific worldview are incompatible. Her argument was supposed to show that commitment to the theory of evolution is incompatible with moral realism. I attempted to turn the anti-realist's argument to my side. I argued that, far from posing a threat to moral realism, evolutionary considerations provide further reason to accept the realist's thesis. My aim in this chapter will be similar. I will consider a familiar argument against moral realism: the argument from disagreement. I will argue that, far from providing a reason to doubt the veracity of moral realism, the existence of moral disagreement counts in favor of the ethical non-naturalist's theses.

The argument from intractable moral disagreement

The argument from disagreement comes in a variety of forms. One can distinguish between formulations of the argument based on the kind of moral disagreement the argument draws on. A popular strand of argumentation in the contemporary literature argues that, if there is intractable moral disagreement, then moral realism is false. There are a variety of ways philosophers have pushed the argument. It has been argued that the existence of intractable moral disagreement gives us reason to doubt that there are any mind-independent moral facts. Alternatively, it has been argued that, if there is intractable moral disagreement, moral knowledge is impossible. While this latter claim is not an immediate problem for moral realism as I have defined it (see chapter one), it is nonetheless a very worrisome conclusion. It is difficult to see the draw of moral realism if the acceptance of moral realism further commits one to the universal denial of moral knowledge.

The contemporary debate over moral disagreement by and large focuses on intractable, as opposed to actual, moral disagreement. There are a variety of ways in which authors have attempted to draw the distinction between actual and intractable moral disagreement. In what follows, little will turn on how one understands this distinction. In all cases, the difference revolves around the conditions under which we can expect moral disagreement to persist. Thus, intractable moral disagreement exists just in case moral disagreement would persist under various counter-factual conditions. Stich and Doris take intractable moral disagreement to consist of moral disagreement that would be present were all parties to the disagreement:

(1) Impartial,

(2) Fully and vividly aware of all relevant non-moral facts,

(3) Free from abnormality. (Doris and Stich)

The impartiality condition requires that parties to the disagreement have no dog in the hunt; intractable moral disagreement requires that no party to the disagreement has a vested (non-moral) interest in the question at hand. Suppose that Sarah's dearly beloved husband, Sam, is on death's row. Sarah is not an impartial party to a disagreement over the moral permissibility of capital punishment.

Stich and Doris further demand that parties to intractable moral disagreement be fully and vividly aware of all relevant non-moral facts. Consider the disagreement over the moral impermissibility of capital punishment. One can often find proponents of capital punishment arguing that the death penalty has a deterrent effect; though it is unfortunate that the death penalty leads to the death of convicted criminals, on the whole, the death penalty saves lives. Opponents of capital punishment draw into doubt these kinds of claims regarding deterrence. Suppose that Lynn and Patrick disagree about the moral permissibility of capital punishment. Furthermore, suppose that this disagreement would be settled were Lynn and Patrick to agree on the deterrent effect of the death penalty. Under such circumstances, it seems clear that Lynn and Patrick's disagreement isn't properly speaking a moral disagreement. Rather, the disagreement is primarily over a non-moral question—it just so happens that the moral question hangs on how one answers this non-moral question. In requiring that parties to intractable moral disagreement be fully and vividly aware of all relevant non-moral facts, Stich and Doris rule out this kind of derivative "moral" disagreement.

Finally, Stich and Doris introduce a catch-all clause: parties to the debate must be free from abnormality. "Freedom from abnormality" is supposed to make moral realism immune from worries that might arise as a consequence of disagreement between individuals with cognitive disabilities. In the words of Stich and Doris: "Obviously, disagreement stemming from cognitive impairments is no embarrassment for moral realism; at the limit, that a disagreement persists when some or all disputing parties are quite insane shows nothing deep about morality" (Doris and Stich 137).

One might demand that disagreement meet some further conditions in order to count as intractable. Notably, Stich and Doris leave out any counterfactual conditions that demand rational debate. It may be that, though disagreement would persist amongst impartial, fully and vividly informed, normal disputants, the disagreement would dissolve after each disputant gives careful consideration to the others' arguments.[41]

Moral realists often respond to worries arising as a consequence of moral disagreement by attempting to explain away the moral disagreement. Each of the above counterfactual conditions highlights one way in which a moral realist might attempt to

[41] As will become apparent later in this chapter, if the argument from intractable moral disagreement is going to be a cause of concern for the moral realist, it must not also be problematic for the scientific realist. Whatever definition one gives of "intractable disagreement," it needs to rule out the possibility of intractable scientific disagreement. Any such definition will need to include significantly more clauses than the four presented here. Minimally, we will want to include some further clause about the parties to the debate having the appropriate competencies. It would be absurd to think that the existence of disagreement regarding the veracity of the Special Theory of Relativity has anything to tell us about physics if one party to the disagreement has never taken a course in mathematics.

show that moral disagreement fails to demonstrate something deep about the nature of morality. The disagreement may merely be the result of partiality, ignorance regarding non-moral facts, "insanity," or the failure to be appropriately reflective regarding one's moral beliefs. If the realist can explain away moral disagreement in any of these ways, it might seem that moral disagreement does not pose a substantial challenge to realism.

As noted earlier, I think that, at least for my purposes here, very little turns on how one chooses to understand "intractable moral disagreement." The reason that this is the case will quickly become apparent. Whatever counterfactual conditions we end up deciding to include in our definition of "intractable moral disagreement," the anti-realist's argument takes one of the following two forms:

(1a) If intractable moral disagreement is possible, then there are no mind-independent moral facts.

(2a) Intractable moral disagreement is possible.

(3a) Therefore, there are no mind-independent moral facts.

Alternatively, sometimes the anti-realist formulates the argument as follows:

(1b) If intractable moral disagreement is possible, then we cannot have moral knowledge.

(2b) Intractable moral disagreement is possible.

(3b) Therefore, we cannot have moral knowledge.

Obviously, the challenge for such arguments is to justify the first and second premises. Why would we think that the possibility of intractable moral disagreement has anything to tell us about the ontological nature of moral facts or the status of moral knowledge? Furthermore, what reason might we have for thinking that intractable moral disagreement is possible? Depending on which author one reads, one will find different answers to these questions. Instead of considering the various formulations of the argument from the possibility of intractable disagreement, I will offer a reason to think that no such argument will be successful.

At least since the early-modern period, moral realists have relied on a “partners in guilt” strategy to defend their view. The strategy is to show that some criticism of moral realism overgeneralizes. Many early modern realists were moral intuitionists; they thought that, via reason, we could have direct access to moral facts. Opponents of moral realism argued that it was implausible to suppose that we could have this kind of direct access to necessary facts. The moral realist responded by drawing an analogy to mathematical knowledge. It seems plausible to suppose that we have knowledge of mathematical facts via reason. It is difficult to imagine how one would tell a story about the genesis of our mathematical knowledge that did not involve the direct apprehension of necessary facts. This allows the intuitionist to argue as follows: if the anti-realist’s argument against moral intuition is successful, we have reason to believe that we have no mathematical knowledge.

This response on behalf of the intuitionist is not, of course, enough to end the debate. It is still open to the anti-realist to either (1) deny mathematical knowledge or (2) attempt to offer some other account of the genesis of our mathematical knowledge; however, as a first blush, the strategy is quite effective. Neither of the anti-realist’s remaining options looks particularly enticing.

Realists have pushed the same style of argument in response to worries arising from moral disagreement. There is widespread disagreement on a variety of topics. We do not think that the fact that there is disagreement about, e.g., the existence of global warming, constitutes a reason to think that (a) there are no mind-independent facts about global warming or (b) that we cannot have knowledge about global warming. Of course, here, the definition of ”intractable moral disagreement” can do a significant amount of work for the anti-realist. It may be implausible to suppose that there is intractable disagreement regarding the existence of global warming whereas it may be plausible to suppose that there is intractable disagreement regarding the existence of moral disagreement.

The version of ethical non-naturalism I have developed offers the ethical non-naturalist novel resources for responding to the argument from intractable moral disagreement. Much like my philosophical predecessors, I will rely on the "partners in guilt" strategy. In any sense in which intractable disagreement is a problem for the version of ethical non-naturalism I have been defending, it is equally a problem for realism about science.

The trick for the proponent of the argument from intractable disagreement will be to show that there's a sense in which the second premise of her argument is true with regard to moral disagreement but not with regard to disagreement regarding scientific questions. Put another way, the proponent of the argument from intractable disagreement needs to show that we have reason to believe the first, but not the second, of the following two claims:

1. Intractable moral disagreement is possible.
2. Intractable scientific disagreement is possible.

If on every understanding of "intractable disagreement" whereby (1) is true, (2) is also true, then the argument from intractable disagreement will apply equally well to scientific realism as it does to moral realism.

At this point, it will be helpful to offer a very brief summary of the argument I have offered in defense of ethical non-naturalism. I first laid out an argument schema that allowed us to identify properties via the predictive power gained by the comprehending application of predicates. I then sketched ways in which we could further characterize these properties by relying on facts about the language we use to describe them. I argued that the moral facts are those facts that we need to posit to explain the deliverances of our moral faculty after factoring out the deliverances of our moral faculty that can be attributed to responsiveness to non-moral non-natural normative properties. The upshot is that we have good reason to believe that an empirical investigation into substantive normative ethical questions could be successful.

A posteriori investigation into non-moral non-natural normative properties would allow us to factor out the influence of these properties on our moral judgments. This *a posteriori* investigation into non-moral non-natural normative properties would presumably proceed via research into the predictive power gained by comprehending application of predicates. Once we have successfully characterized non-moral non-natural normative properties, we can determine which of our moral judgments are a consequence of non-moral non-normative properties—the remaining moral judgments will be a consequence of our awareness of moral properties. Thus, by factoring out the influence of non-moral non-natural normative properties and relying on our moral faculty as a detector, we can expect to be able to empirically investigate normative ethics.

The upshot for the argument from intractable disagreement ought to be clear. In providing a defense of the claim that intractable moral disagreement is possible, the anti-realist must give some reason to believe that intractable moral disagreement is possible that does not also give us some reason to believe that intractable scientific disagreement is possible. For if the anti-realist's argument gives us reason to believe that intractable scientific disagreement is possible then the argument from intractable disagreement overgeneralizes and the ethical non-naturalist can successfully offer a "partners in guilt" defense.

I have just offered a rough sketch of how one might go about empirically investigating normative ethical questions. The possibility of such an investigation brings normative ethics under the broad purview of scientific inquiry. It follows that there can be no sense in which it is true that intractable moral disagreement is possible without it also being the case that it is true that intractable scientific disagreement is possible. If the methods of the sciences rule out intractable disagreement, then the possibility of empirical investigation into normative questions rules out intractable moral disagreement. Alternatively, the methods of the sciences may not rule out intractable disagreement. If this is the case, then scientific realism is as vulnerable to worries stemming from the possibility of intractable disagreement as is moral realism. While it is open to the moral anti-realist to reject scientific realism, at least for the purposes of my project here, such a move is dialectically inadmissible. My project in the dissertation is to assume the truth of the scientific worldview and to see if one can ground a defense of ethical non-naturalism in this assumption. Consequently, for the purposes of this project, showing that an argument against moral realism also undermines scientific realism constitutes a *reductio ad absurdum* of the argument. If the methods of the sciences do not rule out intractable moral disagreement, then, at least given the contours of my project, intractable moral disagreement in some domain ought not be taken to pose a challenge for realism about that domain.

The argument from actual moral disagreement

The objection developed

In many ways, the argument from actual moral disagreement is significantly more challenging for the version of ethical non-naturalism I have developed than is the argument from intractable moral disagreement. Re-consider the hypothesis that underlies the defense of ethical non-naturalism I have offered:

> [H] The function of the mechanisms that ground our dispositions to make moral judgments is to recognize non-natural normative properties.

I have offered two distinct reasons to think that this is the case. In the fifth chapter I noted that our moral judgments are responsive to the presence of role properties and the properties good for and bad for. This is predicted by [H] and consequently gives us some reason to believe [H]. In the sixth chapter I strengthened the case for [H] by arguing that [H] offered the best explanation of the evolution of our moral faculty.

On face, moral disagreement is problematic for [H]. [H] commits us to the view that our dispositions to make moral judgments are reliably responsive to the presence of non-natural normative properties. But actual moral disagreement appears to cast [H] into doubt. If our moral judgments are responsive to the presence of non-natural normative properties, why is there so much moral disagreement?

The argument from actual moral disagreement generally takes the form of an IBE. In his seminal presentation of the argument from actual disagreement, Mackie nicely sketches the anti-realist's argumentative strategy:

> Disagreement about moral codes seems to reflect people's adherence to and participation in different ways of life. The causal connection seems to be mainly that way round: it is that people approve of monogamy because they participate in a monogamous way of life rather than that they participate in a monogamous way of life because they approve of monogamy (Mackie 18).

We can ask this question: what best explains actual moral disagreement? Mackie suggests that the best explanation of moral disagreement is that people's moral judgments are responsive, not to moral properties, but to social conventions. This formulation of the argument from disagreement is not immediately threatening to moral realism. The conclusion of the argument is not a claim about the non-existence of mind-independent moral properties; rather, the conclusion is a claim about what influences our moral judgments. Nonetheless, with a little work, one may be able to offer premises that link this IBE to moral realism. I will not bother to suggest and evaluate any such premises. Even without additional premises, the argument from actual disagreement poses an immediate threat to the defense of ethical non-naturalism I have offered. My defense of ethical non-naturalism requires that we have good reason to accept [H]. Actual moral disagreement appears to give us prima facie reason to reject [H]. If the function of our moral faculty is to recognize non-natural normative properties, and it does so reliably, why do we see so much moral disagreement?

The task for the ethical non-naturalist is to show that, despite apparently substantial actual moral disagreement, we have reason to believe that our moral judgments are reliable responses to non-natural normative properties. The challenge can be sharpened. The moral realist is committed to thinking that, in cases of moral disagreement, at least one party to the disagreement is mistaken. If moral claims are made true by mind-independent properties (that is, if the truth of moral claims is not relative), then if one speaker asserts a moral claim and another speaker asserts the claim's negation, one of the speakers must have said something false. However, the role [H] has played in my defense of ethical non-naturalism commits me to the view that our moral judgments are the consequence of the recognition of non-natural normative facts. It may appear that I find myself in between a rock and a hard place. On one hand, I must hold that, in cases of substantive moral disagreement, one party to the disagreement is wrong. On the other hand, I must hold that each party to the disagreement is making her moral judgment as a response to the recognition of a non-natural normative fact.[42] Presuming that the non-natural normative facts are invariant, it may seem unclear how each party to the disagreement could be responding to a non-natural normative fact. After all, if they are both reliably responding to the same non-natural normative fact, wouldn't we expect them to be making the same moral judgment? Alternatively, if I hold that non-natural normative facts are variant, it seems likely that I have given up my moral realism. Things do not look good for my particular breed of ethical non-naturalism.

Despite this rather bleak perspective, not only do I think that actual moral disagreement fails to constitute a challenge to my brand of ethical non-naturalism, I think that actual moral disagreement offers independent support to the version of ethical non-naturalism I have been defending.

[42] This way of putting the point likely makes my burden sound more significant than it actually is. The best explanation of any number of moral disagreements may be that a party to the debate is particularly unreflective, is mouthing off for attention, etc. Nonetheless, at least for the time being, I will accept this unreasonable burden. The aim is to see how successful [H] can be in providing an explanation of actual moral disagreement. The more disagreement [H] can explain, the more consilient [H] looks. If we want to see how much disagreement [H] can explain, we should start by attempting to explain *all* moral disagreement in terms of [H]. Other explanations of particular instances of disagreement can always be added back into our explanatory story.

A response to the argument from actual moral disagreement

In the previous chapter I developed an etiological account of the development of our moral faculty. According to the picture I sketched, the moral faculty evolved to recognize non-natural normative properties. In particular, evolutionary pressures pushed us in the direction of recognizing role properties and the properties of being good for some organism and being bad for some organism. On this view, our ability to recognize non-natural moral properties is a spandrel. Given that our moral faculty evolved to respond to non-natural normative properties, as opposed to some particular subset thereof, I predicted that the presence of one non-natural normative property would influence our judgments about the presence of moral properties.

This etiological hypothesis can be nicely expanded to explain the existence of actual moral disagreement. However, before engaging in this explanatory task, it will be helpful to have an instance of moral disagreement on the table around which we can structure the subsequent discussion. The following quotation illustrates a striking variation in cultural practice—a variation that we can expect would give rise to moral disagreement:

> [T]he Yanomamo of Brazil, or at least Yanomamo men, provide an example of a group whose moral standards seem to contrast strikingly with ours. Central to norms for the behavior of Yanomamo men is the high value placed on "fierceness": in warfare, toward other men in one's community, and toward women--especially, one's wives. The ability to intimidate one's peers through violence and threats of violence is highly regarded. Severe violence towards one's wives is an esteemed method for demonstrating one's fierceness—it is deemed appropriate for a man to hurt his wife severely, e.g., by shooting her in the leg with an arrow, for such reasons as being too slow in serving him dinner. Warfare between tribes occurs frequently with little or no provocation; killing the men and capturing the women of neighboring tribes is seen as a highly desirable end, and treacherous breaking of alliances is highly regarded as a means to this end... [It is typical for] members of a neighboring village... [to be] invited to share in a feast, ostensibly to make peace, only to be subjected to a surprise attack in which the hosts aim to kill as many visiting men and capture as many women as possible—successful planning and execution of such a treacherous attack is considered highly admirable. (Bennigson 412)

Character traits and actions that we would consider morally depraved—deplorable—are considered by male Yanomami to be worthy of high commendation. It seems clear that there is significant disagreement between male Yanomami and ourselves regarding moral norms.

Can the etiological hypothesis I developed in the previous chapter help us make sense of this instance of disagreement? In order for the etiological hypothesis to be helpful, I must identify some non-natural non-moral normative property that could plausibly be responsible for skewing either the moral judgments of the Yanomamo or our moral judgments. In the previous chapter I argued that the property of being bad for humans can explain why poison disgust can influence our moral judgments. In the case of moral disagreement, the other kind of non-natural normative property we've identified—role properties—can do valuable explanatory work.

A brief review of role properties is in order. Various social roles exist in order to fulfill various functions. Thus, the role of the teacher is to educate. The role of the bus driver is to drive a bus. The role of the chef is to cook food. Each role can be performed in better and worse ways. Thus, a teacher can be good at educating or bad at educating. A chef can cook well or cook poorly. The teacher who is good at educating is a good teacher. The teacher who is bad at educating is a bad teacher. The same goes for the other roles. The chef who cooks well is a good chef. The chef who cooks poorly is a bad chef.

The variation in possible types of social roles appears to be limited only by the activities that humans can engage in. For any activity that can be done well, or done poorly, there can be a concomitant social role. Furthermore, "activity" should be broadly construed. The good father is the father who raises his children well. Raising one's children only counts as an activity when "activity" is broadly construed. The limits on possible social roles are broad.

In light of the broad range of possible social roles, we can expect social roles to vary culturally. Thus, while it was common to have someone play the role of matchmaker in medieval shtetls, it is rare to find someone playing this role in contemporary America. Perhaps more relevant for our discussion, in Nazi Germany one could find individuals filling the role of concentration camp commandant. Happily, one is unlikely to find a person fulfilling this role in modern day Europe. Importantly, just as it is possible to be a good father, a good chef, and a good teacher, it is also possible to be a good concentration camp commandant.

We are now well positioned to reconsider the moral disagreement that appears to exist between ourselves and the Yanomami. Whereas the Yanomami think that violence is praiseworthy, we consider it morally deplorable. Some explanation of this disagreement is owed and, as I understand my burden, I am committed to (1) holding that at least one party to the debate is wrong and (2) that both our moral judgment and the moral judgment of the Yanomami are a response to non-natural normative properties.

It is widely accepted that gender is a social construct (Freud). One social role a person may fill is the role of *being a man*. (Importantly, as I understand role properties, they vary enormously per society. It is perhaps misleading to talk about there being a single role property: *being a good man*. More accurately, there are a range of distinct role properties [$\textit{good man}_{\textit{Midwest American}}$, $\textit{good man}_{\text{Yanomami,}}$$\textit{good man}_{\text{Hasidic Jew}}$, ..., $\textit{good man}_{\text{n}}$] which it can be helpful to think about in terms of the nomenclature of "being a good man.") What constitutes successfully fulfilling this social role will vary per culture—different cultures will have different standards of masculinity. For the Yanomami, one successfully fulfills the role of being a man only if one is "fierce." Thus, while shooting one's wife in the leg with an arrow is morally impermissible, it may also be partly constitutive of being a good man in Yanomami culture. We can explain the disagreement between the Yanomami and ourselves by pointing to distinct non-natural normative properties each party to the dispute is responding to. While we may be responding to the *prima facie* wrongness of harming another, the Yanomami are likely responding to the non-natural normative property of being a good man. In the previous chapter I presented an etiological account of our moral faculty that had the upshot that we should expect our moral judgments to be influenced by the presence of non-moral non-natural normative properties. Not only does this account help explain the evolution of morality, it nicely positions us to explain the existence of moral disagreement.

Before proceeding to potentially more problematic instances of apparent moral disagreement, by way of illustrating the explanatory power of the current proposal, let us consider a related example:

> In 2006, Zahra al-Azzo was kidnapped and raped near her home in Damascus, Syria. Following her safe return, her older brother stabbed and murdered her in her sleep. In response to his killing her, Zahra's family held a large celebration that night. According to the United Nations Population Fund, 5,000 similar "honor killings" occur each year. (Zoepf)

Here we have another striking example of moral disagreement. We are of the opinion that the killing of Zahra was morally unjustifiable—murder. Zahra's family saw it as called for; an event worthy of celebration. The etiological hypothesis I developed in the sixth chapter once again offers to provide an elegant explanation of the apparent disagreement. In the relevant Syrian subculture, protecting a family's honor is partially constitutive of being a good son. As before, the etiological hypothesis developed in the sixth chapter offers to explain the existence of moral disagreement. At least in the cases we have considered, moral disagreement exists because our moral faculty evolved to recognize non-natural normative properties, generally, and not moral properties specifically. Consequently, the presence of non-moral non-natural normative properties can influence our moral judgments.

The argument from actual moral disagreement turned

How does this explanation of the existence of moral disagreement fare in competition with other explanations? I think the answer is: very well. Let us briefly reconsider Mackie's point about the existence of moral disagreement:

> Disagreement about moral codes seems to reflect people's adherence to and participation in different ways of life. The causal connection seems to be mainly that way round: it is that people approve of monogamy because they participate in a monogamous way of life rather than that they participate in a monogamous way of life because they approve of monogamy. (Mackie 18)

It seems that any account of moral disagreement will require some explanation of why some of the most striking moral disagreement appears to fall along cultural fault lines. Let us consider a promising anti-realist explanatory strategy.

Perhaps the most promising anti-realist friendly explanatory approach is to hold that moral disagreement co-varies with the approval and disapproval of members of one's culture. Thus, one could attempt to explain the disagreement between ourselves and the Yanomami in terms of the attitudes prevalent in each culture. Individual Yanomamo judge violence permissible because there is a prevailing attitude of approval towards violence in the Yanomami culture. In contrast, we judge that violence is *(prima facie)* impermissible because, in our society, there is a prevailing attitude of disapproval towards violence.

The first thing to note that this explanatory strategy is not, as it stands, incompatible with the non-naturalist friendly explanatory strategy I have thus far offered. Role properties are properties had in virtue of occupying a social role. There is little doubt that social roles are, at least partly, constituted by the attitudes of the members of a society. Thus, part of what, e.g., makes Sam a teacher, are the attitudes others have towards Sam. The non-naturalist can happily accept that moral judgments vary along cultural boundaries because attitudes of approval and disapproval vary amongst cultures. The non-naturalist need only add one further caveat: the reason moral judgments vary alongside attitudes of approval and disapproval is because these attitudes are partly constitutive of facts regarding social roles.

The same argumentative strategy can be used on behalf of the ethical non-naturalist with regard to any attempt to demonstrate that moral disagreement is best explained via cultural variance in the prevalence of certain mental states. Whatever mental state the anti-realist points to in order to explain inter-cultural moral disagreement, it must be the case that the mental state in question is responsible for societal expectations regarding appropriate behavior. Given that social roles are socially constructed, it is always open to the ethical non-naturalist to hold that the mental states in question are partly constitutive of a social role and, consequently, moral disagreement is fundamentally caused, not by moral judgments being responsive to prevalent mental states in one's culture, but by moral judgments being responsive to non-natural normative properties associated with the social roles which are constituted by the mental states in question.

The power of this response on behalf of the ethical non-naturalist ought not be underestimated. The argument from actual disagreement was supposed to take the following form: the best explanation of actual moral disagreement is that our moral judgments are responsive to mind-dependent, not mind-independent, properties. It may be helpful to put the argument schematically: the best explanation of some instance of actual moral disagreement, D, is that our moral judgments, J, are responsive to mind-dependent properties, MD, and not mind-independent properties, MI. The version of ethical non-naturalism I have developed allows the non-naturalist to co-opt any putative explanation the anti-realist offers. For any MD the anti-realist identifies, so long as it is plausible to hold that MD is partially constitutive of social roles, the non-naturalist can hold that the best explanation of D is that our moral judgments are responsive to some MI corresponding to the social role partially constituted by MD.

Importantly, one might think that this explanatory strategy on the part of the ethical non-naturalist is ontologically promiscuous. While it is true that the ethical non-naturalist can attempt to explain actual moral disagreement this way, we shouldn't think that any such explanation is a best explanation. After all, for the explanation to work, we are forced to posit the existence of non-natural normative properties over-and-above the mind-dependent properties both the anti-realist and the ethical non-naturalist are committed to. We should, however, not be particularly impressed with this rejoinder on the part of the anti-realist. If my arguments in the fourth and fifth chapter of the dissertation are successful, we have independent reason to posit the existence of role properties *qua* non-natural normative properties. Consequently, the non-naturalist friendly approach to explaining actual moral disagreement sketched above comes at no extra ontological cost.

I opened this chapter by suggesting that my defense of ethical non-naturalism was faced with the following problem: actual moral disagreement casts doubt on the hypothesis that the function of the mechanisms that ground our dispositions to make moral judgments is to recognize non-natural normative properties. If our moral faculty tracks invariant moral facts it may be difficult to see how there could be moral disagreement. I take myself to have now offered a defense on behalf of the ethical non-naturalist to this charge. The account of non-naturalism I have developed appears to have the resources to explain actual disagreement without having to surrender the thesis that the function of our moral faculty is to recognize non-natural normative properties.

My aim in this chapter is, however, significantly more ambitious. I am not content merely to show that the version of ethical non-naturalism I have defended has the resources to respond to worries that might stem from disagreement. I would like to further demonstrate that the version of ethical non-naturalism I have developed offers the best explanation of moral disagreement. Success in this task requires that I show that the above sketched explanatory strategy on behalf of the ethical non-naturalist offers the best explanation of actual moral disagreement. Is there any reason to prefer an account of moral disagreement whereby moral disagreement is a consequence of responsiveness to non-moral non-natural normative properties? I think that the answer is "yes." To show that this is the case, we first need to get the relevant evidence on the table.

As already noted, there appears to be substantial actual moral disagreement. This is not, however, to say that there are no universal moral norms: "prohibitions of murder, rape and other types of aggression appear to be universal..." (Mikhail 143). Though one can find significant moral disagreement, on at least some questions, it appears that there is cross-cultural moral agreement. Any explanation of moral disagreement must be capable of making sense of both widespread moral disagreement and widespread moral agreement.

Furthermore, given the assumption of weak moral nativism made in the previous chapter, we are committed to offering an account of our moral faculty whereby the moral faculty is responsive to some set of properties found in the external world. Our moral judgments can only be fitness conducive if they reliably change our behavior with regard to some set of stimulus.

Additionally, as noted in the fifth chapter, it appears that our moral judgments are responsive to non-moral non-natural normative properties. Consequently, any account of our moral faculty must be in a position to explain how it is that our moral judgments are responsive to non-moral non-natural normative properties while also accounting for the phenomenon of moral disagreement.

Finally, as noted in the previous chapter, poison disgust influences our moral judgments. Any account of the function of our moral faculty must be able to account for the influence of poison disgust on our moral judgments.

The task then is to give an account of our moral faculty that can: (1) explain the existence of widespread moral disagreement, (2) explain the existence of widespread moral agreement, (3) identify some set of external world properties our moral judgments are responsive to, and (4) explain the influence of poison disgust on our moral judgments. Note how nicely [H] fares on each account. Moral disagreement can be explained via the influence of non-moral non-natural normative properties on our moral judgments. Moral agreement can be explained by the fact that our moral faculty is responsive to mind-independent moral facts. Non-natural normative facts are found in the external world. Finally, poison disgust is a response to instantiations of the non-natural normative property *bad for humans* and, as such, we would expect poison disgust to influence our moral judgments. [H] is strikingly consilient—capable of explaining an enormous variety of apparently unrelated phenomena.

There is reason to think that the explanation of moral disagreement [H] provides will be better than any explanation the anti-realist can offer. For the anti-realist's explanation to be equally consilient, she must identify a property that is either causally responsible or partially constitutive of each of the above phenomena and is not a non-natural mind-independent normative property. The phenomena in need of explanation are, on their face, strikingly disparate. It is difficult to imagine any anti-realist friendly property that might serve to offer a unifying explanation of moral disagreement, moral agreement, and the influence of poison disgust on our moral judgments.

My considered defense of ethical non-naturalism required that we have good reason to accept [H]. It appears that we now have some further reason to believe that [H] plays a role in a best explanation. [H] is explanatorily powerful, allowing us to offer a unified explanation of phenomena that previously appeared unrelated. We now added, to the list of phenomena [H] can explain, the existence of moral disagreement.

Two further cases of disagreement considered

Before concluding this chapter, I would like consider two further cases of actual moral disagreement. In some ways, though the instances of moral disagreement I have thus far offered are striking, they are relatively easily dealt with by the form of ethical non-naturalism I have developed. Other instances of moral disagreement, though perhaps less intuitively problematic to moral realism, pose a more substantial threat to the version of ethical non-naturalism on offer.

Reconsider a case presented earlier in this chapter. Sarah is opposed to the death penalty and the best explanation of Sarah's opposition to the death penalty is that her husband, Sam, is on death row. This kind of explanation has, historically, been used by the moral realist to push back against the argument from actual disagreement. The thought was that, if actual moral disagreement could be shown to be a consequence of psychological bias, actual moral disagreement would fail to demonstrate anything substantive about moral properties. This realist mainstay in response to the argument from actual disagreement, however, poses a potential problem for the defense of ethical non-naturalism I have offered. I have argued that the function of our moral faculty is to recognize non-natural normative properties; however, it would appear that the best explanation of Sarah's belief that capital punishment is morally impermissible has less to do with non-natural normative facts and more to do with psychological bias.

There is nothing immediately problematic for my account with explaining Sarah's moral opposition to capital punishment in terms of psychological bias. Psychological bias appears to infect a wide variety of judgment making processes. We do not think, in other instances where bias can affect our judgments, that the existence of bias counts as evidence regarding the role of our cognitive capacities. Consequently, we ought not take bias to pose any particular problem here.

Nonetheless, wherever possible, I would like to rely on [H] to explain moral disagreement. Suppose that the best explanation of Sarah's judgment that capital punishment is morally impermissible is that Sarah's husband is on death row. Can we subsume this explanation under the explanatory rubric offered by [H]? Once again, I think that the answer is "yes."

Why do male members of the Yanomami judge it morally permissible to cause serious harm to their wives? The moral faculty of the Yanomami are picking up on a non-natural normative property—the role property of being a good man. But note the relevant difference between this case and the case of Sarah. The Yanomani man who judges that it is morally permissible to cause serious harm to his wife is making a judgment about the permissible behavior of a Yanomami man. There are both moral and non-moral norms that govern the behavior of a Yanomami man; if [H] is true it comes as little surprise that conflicting non-natural norms regarding the permissibility of behavior could skew one's moral judgments. The story is not nearly so straightforward with Sarah. Sarah's judgment is about the moral permissibility of capital punishment, not about the moral permissibility of any action she might undertake. It may be partially constitutive of being a good partner that one be opposed to the execution of one's loved one. This role property puts a norm on the appropriate behavior of a partner. On the model of the response we developed, the relevant role property would explain Sarah's judgment that it is morally obligatory for Sarah to be opposed to the execution of Sam. If the role property of being a good partner is going to explain Sarah's judgment that capital punishment is morally impermissible we need to provide some further link between the role property, which establishes norms about how Sarah ought to behave, and Sarah's judgment, which is about how others ought to behave (on the presumption that Sarah is not, e.g., a judge or an executioner).

The missing link is our penchant for internalizing norms. A norm is internalized when "[b]ehaviors that are dictated by social norms become the ends that individuals desire" (Horne 336). When Sarah internalizes the norms constituted by the non-natural property of being a good partner, she comes to be intrinsically motivated to act in ways consistent with the norm. Thus, Sarah's judgment that capital punishment is morally impermissible can be explained by the concatenation of the fact that she has internalized the relevant norms and the presence of the appropriate non-natural normative property.

Had she not internalized the norms, she would not have judged that capital punishment is morally impermissible. Alternatively, had it not been the case that being a good partner committed Sarah to being opposed the execution of Sam, Sarah would not have judged that capital punishment is morally impermissible (assuming, of course, that her judgment was not causally overdetermined). With the addition of an auxiliary hypothesis (to which we were all already committed), [H] can explain Sarah's judgment that capital punishment is morally impermissible.

One final case of moral disagreement, on offer from Brandt, demands consideration:

> [Hopi c]hildren sometimes catch birds and make "pets" of them. They may be tied to a string, to be taken out and "played" with. This play is rough, and birds seldom survive long. [According to one informant:] "Sometimes they get tired and die. Nobody objects to this (as cited in Doris and Stich 130).

I presume that it is morally impermissible to torture animals to death as part of "play." If I am correct in this moral assumption, some explanation is owed: why do members of some cultures fail to recognize this moral fact?

This case is particularly problematic for the ethical non-naturalist. The crux of the problem is that birds are not humans. The account of moral disagreement I have relied on makes heavy weather of the way in which role properties can distort our moral judgments. Role properties, arising from the structure of social roles, constitute sets of norms regarding the appropriate behavior of the members of a culture. Thus, the role property of being a good man determined the "appropriate" behavior towards a woman on behalf of a Yanomami man. The role property of being a good son determined the "appropriate" behavior of Zahra's brother towards Zahra. In each case, the role property appears to be relational, i.e. the role property holds with regard to any two individuals who fulfill the relevant social role: ${}_{\text{Yanomami man}}R_{\text{woman}}$ and ${}_{\text{brother}}R_{\text{sister}}$. The same model does not appear to apply to birds. It is plausible to think that birds do not fill a social role.

There is a tempting solution to the putative problem. Being a good teacher may be a property one can only have in virtue of one's relationship with others who fill a social role (i.e. one's students); however, being a good welder surely does not require that one stand in any particular social relationships. It need merely be the case that one welds, and that one welds well. Being a good welder may be a relational property—one must stand in the right kind of relationship with the artifacts one has welded; however, it does not look like a relational property that is determined in virtue of the welder's relationship to individuals who fill other social roles.

One might be tempted to try and find a property analogous to good welder to explain why the Hopi think that it is morally permissible for their children to torture birds to death. It seems unlikely that any such explanation will be forthcoming. The trouble is that, while it is very plausible to think that being a good welder ought to be understood in terms of a welder's relationship to the artifacts she creates, it is implausible to suppose that the role of good child ought to be understood in terms of a child's relationship with a bird. Of course, nothing forces us to make sense of the moral disagreement in question in terms of the role property *good child*; nonetheless, the point stands. It seems rather far-fetched to suppose that Hopi children occupy some social role that ought to be understood in terms of a relationship with a bird.

I think that this apparent problem for the non-naturalist explanation of moral disagreement only arises because we are thinking of the social roles a bird might fill too narrowly. Perhaps surprisingly, just as we can explain our disagreement with Yanomami men by considering how the role Yanomami man is partially constituted by a relationship with a different social role, woman, we can explain our disagreement with the Hopi by pointing to a social role children occupy where this social role is partially constituted by a relationship with a social role that is filled by birds. The trick, of course, is to find some social role that it is plausible to think that a bird could fill.

It is a widely accepted sociological fact that humans tend to divide the world into, broadly speaking, two kinds of people. Some people are members of one's in-group and others are members of one's out-group. Members of the ingroup are people like oneself; that is, members of the ingroup are people who are part of a group with which one identifies oneself. Family members, members of the same tribe, members of the same religion, and people who share our ethnicity often count as members of the ingroup. Members of the outgroup are people who are not like us; that is, members of the outgroup are people who are not a part of a group with which one identifies oneself. People who are not family members, not of our tribe, not of our religion, and not of our ethnicity are often members of the outgroup. It is a further widely accepted sociological fact that we tend to treat members of the outgroup poorly. Actions that would be morally impermissible if taken towards members of the ingroup are often considered morally permissible if aimed at members of the outgroup. Part of what it is to be a member of the ingroup is to be worthy of moral consideration. In contrast, part of what it is to be a member of the outgroup is to be unworthy of moral consideration. The good ingroup member dedicates her energies to the ingroup, even when this comes at substantial cost to members of the outgroup.

We are now in a position to explain the moral disagreement that arises between the Hopi and us. The outgroup consists of everything that is not a member of the ingroup. Much too frequently, the outgroup is composed of people who are superficially unlike us: people who speak a different language, have a different skin tone, or have different sexual preferences. Importantly, there is no reason to suppose that an organism must be a person to be a member of the outgroup. That we consider non-human animals to be a part of the outgroup may go a long way towards explaining our willingness to engage in the practices constitutive of factory farming (Plous). It is nearly certain that a bird does not constitute a member of the ingroup for a Hopi child. Consequently, the bird is a member of the outgroup. Membership in the outgroup is grounds for not being given moral consideration. In this instance, the disagreement arises as a consequence of the non-natural normative property of being a good ingroup member. The pleasure the child receives by torturing a bird to death is pleasure felt by a member of the ingroup. The good ingroup member values this pleasure over the pain felt by a member of the outgroup, i.e. the bird. Once again, the non-naturalist finds that, on the assumption that [H] is true, she can explain the existence of actual moral disagreement.

Conclusion

In this chapter I considered two distinct arguments that were supposed to pose a problem for moral realism: the argument from intractable moral disagreement and the argument from actual moral disagreement. The version of ethical non-naturalism I have defended offers to ground an *a posteriori* research program in normative ethics. The upshot is that intractable moral disagreement, if there is any such thing, constitutes a subset of intractable scientific disagreement. If intractable moral disagreement poses a problem for moral realism, it also poses a problem for scientific realism. The argument from intractable moral disagreement overgeneralizes.

The argument from actual moral disagreement poses a significantly more straightforward problem for the defense of ethical non-naturalism I developed. At first glance, it may be difficult to see how, if there is actual moral disagreement, [H] could be true. In the second half of the chapter I argued that, far from counting against [H], the existence of actual moral disagreement constitutes evidence for [H]. Not only is [H] compatible with the existence of moral disagreement, [H] can explain actual disagreement in a way that unifies the existence of actual disagreement with apparently unrelated phenomena, e.g. the existence of widespread moral agreement and the influence of poison disgust on our moral judgments.

CONCLUSION

I have now concluded my methodologically naturalist defense of ethical non-naturalism. In my more optimistic moments, I hope that I have offered a scientifically plausible defense of ethical non-naturalism. In my more pessimistic moments, I fear that my project is an instance of the worst kind of pseudo-science. I suspect that, in reality, the truth lies somewhere in the middle. I will close my dissertation by offering a brief evaluation of the work found in the preceding seven chapters.

If the reader was expecting to find a definite and scientifically rigorous defense of ethical non-naturalism—as I sometimes suggested she would—I suspect the reader is disappointed. I do not take myself to have shown that, if one takes science seriously, one is thereby committed to ethical non-naturalism. More modestly, I hope to have sketched a research program that may, some day down the road, have this consequence. As I see it, there are two primary obstacles to the success of the research program I have sketched. First, more needs to be said in support of the test for abstraction I developed in the fourth chapter. In particular, I would like to be able to offer better evidence that the calculation of the three dimensional structure of, for example, the titin protein requires exponential time. Second, more needs to be said in support of [H]. I have argued that [H] is a shockingly consilient thesis; however, I have looked at a tiny fraction of the phenomena [H] might explain. Further exploration of [H]'s explanatory power is owed. [H] is a scientific thesis. We would do well to look for experimental verification.

Nonetheless, I consider my project successful. I set out to accomplish a task that many consider impossible. Even if I have failed to produce a rationally compelling defense of ethical non-naturalism, it is no small task to have developed the outlines of an empirical research program that, could one day, make ethical non-naturalism a mainstay of the scientific worldview. More importantly, my project should not be judged by the standards generally applied to traditional philosophy. I have not produced deductively sound arguments in support of ethical non-naturalism. But I never set out to do so. Part and parcel of offering a methodologically naturalist defense of ethical non-naturalism is eschewing traditional philosophical standards of argument evaluation in favor of the standards by which scientific theories are judged. The ultimate success of this project will only be determined years down the road. Emerging research programs are not judged on their first-shot success, but rather on the evidence they amass over the course of their development. I take myself to have offered some evidence that commitment to the scientific worldview commits one to ethical non-naturalism. The work that remains, for myself and for others, is to turn the fledgling research program I have sketched here into a full-fledged scientific endeavor.

REFERENCES

Adams, Fred C. "The future history of the universe." *Cosmic update*. Ed. Farzad Nekoogar. New York: Springer, 2012. 71-118.

Armstrong, D.M. *A theory of universals: Universals and scientific realism volume II*. Cambridge University Press, 1978.

Aristotle. *Nicomachean Ethics*. Trans. Hippocrates G. Apostle. Des Moines, Iowa: The Peripatetic Press, 1984.

Ayer, AJ. Language, truth and logic. London: Victor Gollancz Ltd. 1936.

Bekoff, Marc, and Colin Allen. "Teleology, function, design and the evolution of animal behaviour." *Trends in ecology & evolution* 10.6 (1995): 253-255.

Beauchamp, Tom L., and James F. Childress. *Principles of biomedical ethics*. 5th ed. Oxford University Press, 2001.

Bennigson, Thomas. "Irresolvable disagreement and the case against moral realism." *The Southern journal of philosophy* 34.4 (1996): 411-437.

Blackburn, Simon. *Ruling passions*. Oxford: Clarendon Press, 1998.

Blackburn , Simon. "Antirealist expressivism and quasi-realism." *The Oxford Handbook of Ethical Theory*. Ed. David Copp. Oxford: Oxford University Press, 2006. 146-162.

Boyd, Richard N. "On the current status of the issue of scientific realism." *Erkenntnis* 19.1-3 (1983): 45-90.

Brandt, Richard B. *Hopi ethics: A theoretical analysis*. University of Chicago Press, 1974.

Brink, David. *Moral Realism and the Foundations of Ethics*. Cambridge University Press, 1989.

Card, Robert F. "Conscientious objection and emergency contraception." *The American Journal of Bioethics* 7.6 (2007): 8-14.

Cartlidge, Edwin. "Loose cable may unravel faster-than-light result." *Science* 335.6072 (2012): 1027-1027.

Chakravartty, Anjan. "Scientific realism." *Stanford Encyclopedia of Philosophy*. Stanford Encyclopedia of Philosophy, n.d. Web. 11 June 2013.

Chomsky, Noam. "A review of BF Skinner's Verbal Behavior." *Language* 35.1 (1959): 26-58.

Colburn, Timothy, and Gary Shute. "Abstraction in computer science." *Minds and Machines* 17.2 (2007): 169-184.

Conee, Earl, and Richard Feldman. *Evidentialism: Essays in Epistemology*. Clarendon Press, 2004.

Curlin, Farr A. "Caution: Conscience is the limb on which medical ethics sits." *The American Journal of Bioethics* 7.6 (2007): 30-32.

Cushman, Fiery, Liane Young, and Joshua Greene. "Multi-system moral psychology." *The Oxford handbook of moral psychology*. Ed. John M. Doris. New York: Oxford University Press, 2010. 47-71.

Dennett, Daniel C. "Intentional systems." *The Journal of Philosophy* 68.4 (1971): 87-106.

Dennett, Daniel. "Real patterns." *The Journal of Philosophy* 88.1 (1991): 27-51.

Dennett, Daniel. "True believers: The intentional strategy and why it works." *Mind Design II*. Ed. John Haugland. MIT Press, 1997. 57-80.

Machery, Edouard and Ron Mallon. "The evolution of morality." *The moral psychology handbook*. Ed. Doris, John M., and Fiery Cushman. New York: Oxford University Press, 2010. 4-46.

Doris, John M. and Stephen Stich. "As a matter of fact: Empirical perspectives on ethics." *The Oxford handbook of contemporary philosophy*. Ed. Frank Jackson and Michael Smith. New York: Oxford University Press, 2005. 114-152.

Doris, John M. "Persons, situations, and virtue ethics." *Nous* 32.4 (1998): 504-530.

Dretske, Fred. "Epistemic operators." *The Journal of Philosophy* 67.24 (1970): 1007-1023.

Dunbar, Robin. "Why humans aren't just Great Apes." *Issues in Ethnology and Anthropology* 3 (2008): 15-33.

Dwyer, Susan, Bryce Huebner, and Marc D. Hauser. "The linguistic analogy: Motivations, results, and speculations." *Topics in cognitive science* 2.3 (2010): 486-510.

Elwood, Lisa S., and Bunmi O. Olatunji. "A cross-cultural perspective on disgust." *Disgust and Its Disorders: Theory, Assessment, and Treatment Implications*. Ed. Bunmi O. Olatunji and Dean McKay. Washington D.C.: American Psychological Association, 2009. 99-122.

Enoch, David. "The epistemological challenge to metanormative realism: how best to understand it, and how to cope with it." *Philosophical Studies* 148.3 (2010): 413-438.

Enoch, David. *Taking morality seriously: A defense of robust realism*. Oxford University Press, 2011.

Eskine, Kendall J., Natalie A. Kacinik, and Gregory D. Webster. "The Bitter truth about morality: Virtue, not vice, makes a bland beverage taste nice." *PloS ONE* 7.7 (2012): e41159.

Eskine, Kendall J., Natalie A. Kacinik, and Jesse J. Prinz. "A Bad Taste in the Mouth Gustatory Disgust Influences Moral Judgment." *Psychological Science* 22.3 (2011): 295-299.

Fales, Evan. “Naturalist moral realism.” *G-d and morality: four views.* Ed. R. Keith Loftin. Intervarsity Press, 2012.

Fessler, Daniel MT, and Edouard Machery. "Culture and cognition." *Oxford handbook of philosophy and cognitive science.* Ed. Eric Margolis, Richard Samuels, and Stephen P. Stich. New York: Oxford University Press, 2012. 503-527.

Firth, Roderick. "Ethical absolutism and the ideal observer." *Philosophy and Phenomenological Research* 12.3 (1952): 317-345.

Freud, Sophie. "The social construction of gender." *Journal of Adult Development* 1.1 (1994): 37-45.

Fumerton, Richard. "The paradox of analysis." *Philosophy and Phenomenological Research* 43.4 (1983): 477-497.

Gibbard, Allan. *Wise choices, apt feelings: A theory of normative judgment.* Harvard University Press, 1992.

Gibbard, Allan. *Thinking how to live.* Harvard University Press, 2008.

Graber, Abraham. "Medusa’s Gaze Reflected: A Darwinian Dilemma for Anti-Realist Theories of Value." *Ethical Theory and Moral Practice* 15.5 (2012): 589-601.

Greene, Joshua. “The secret joke of Kant’s soul.” *Moral psychology: historical and contemporary readings.* Ed. Thomas Nadelhoffer, Eddy Nahmias, and Shaun Nichols. Wiley-Blackwell, 2010. 359-372.

Haney, Craig, Curtis Banks, and Philip Zimbardo. "Interpersonal dynamics in a simulated prison." *International Journal of Criminology & Penology* 1 (1973): 69-97.

Hare, Richard Mervyn. *Freedom and reason.* Oxford Paperbacks, 2003.

Harman, Gilbert. “Ethics and observation.” *Foundations of ethics: an anthology.* Ed. Russ Shafer-Landau and Terence Cuneo. Malden, MA: Blackwell, 2007. 333-337.

Horne, Christine. "The internal enforcement of norms." *European Sociological Review* 19.4 (2003): 335-343.

Hume, David. *A treatise of human nature.* Ed. L.A. Selby-Bigge. Oxford: Clarendon Press, 1978.

Hursthouse, Rosalind. *On virtue ethics.* New York: Oxford University Press, 1999.

Isen, Alice M. and Paula F. Levin. "Effect of feeling good on helping: cookies and kindness." *Journal of personality and social psychology* 21.3 (1972): 384-388.

Joyce, Richard. *The evolution of morality.* The MIT Press, 2006.

Kelly, Daniel. *Yuck!: the nature and moral significance of disgust.* The MIT Press, 2011.

Krauss, Lawrence and Glenn D. Starkman. “Universal limits on computation.” *arXiv.org.* arXiv:astro-ph/0404510, 2004. Web. 11 June 2013.

Kulstad, Mark and Carlin, Laurence. “Leibniz’s philosophy of mind.” *Stanford Encyclopedia of Philosophy*. Stanford Encyclopedia of Philosophy, 21 September, 2008. Web. 12 June 2013.

Kumar, Victor, and Richmond Campbell. "On the normative significance of experimental moral psychology." *Philosophical Psychology* 25.3 (2012): 311-330.

Ladyman, James, et al. *Every thing must go: Metaphysics naturalized*. Oxford: Oxford University Press, 2007.

Leiter, Brian. "Moral facts and best explanations." *Social Philosophy and Policy* 18.2 (2001): 79-101.

Lloyd, Seth. "Computational capacity of the universe." *Physical Review Letters* 88.23 (2002): 1-4.

Mackie, J.L. “The Subjectivity of Values.” *Foundations of Ethics: An Anthology*. Ed. Russ Shafer-Landau and Terence Cuneo. Malden, MA: Blackwell Publishing Ltd. 13-22.

Mikhail, John. "Universal moral grammar: Theory, evidence and the future." *Trends in cognitive sciences* 11.4 (2007): 143-152.

Milgram, Stanley. "Some conditions of obedience and disobedience to authority." *Human relations* 18.1 (1965): 57-76.

Moore, G. E. *Principia Ethica*. Courier Dover Publications, 2004.

Olatunji, Bunmi O., et al. "Core, animal reminder, and contamination disgust: Three kinds of disgust with distinct personality, behavioral, physiological, and clinical correlates." *Journal of Research in Personality* 42.5 (2008): 1243-1259.

Papineau, David. “Naturalism.” *Stanford Encyclopedia of Philosophy*. Stanford Encyclopedia of Philosophy, n.d. Web. 7 June 2013.

Plantinga, Alvin. *Warrant and proper function*. Oxford University Press, 1993.

Plous, Scott. "Psychological mechanisms in the human use of animals." *Journal of Social Issues* 49.1 (1993): 11-52.

Prinz, Jesse and Shaun Nichols. “Moral emotions.” *The moral psychology handbook*. Ed. John M. Doris. New York: Oxford University Press, 2010. 111-146.

Prinz, Jesse. "The emotional basis of moral judgments." *Philosophical Explorations* 9.1 (2006): 29-43.

Rawls, John. *Justice as fairness: A restatement*. Harvard University Press, 2001.

Rawls, John. "Kantian constructivism in moral theory." The journal of philosophy 77.9 (1980): 515-572.

Ridge, Michael. “Moral Non-Naturalism.” *Stanford Encyclopedia of Philosophy*. Stanford Encyclopedia of Philosophy, n.d. Web. 7 June 2013.

Ross, William David. *The right and the good*. Oxford University Press, 2002.

Rozin, Paul. “The selection of food by rats, humans, and other animals.” *Advances in the study of behavior*. Ed. Jay S. Rosenblatt. New York: Academic Press, Inc., 1976. 21-76.

Schroeder, Mark. *Being for: Evaluating the semantic program of expressivism*. Oxford University Press, 2008.

Schroeder, Mark. "What is the Frege-Geach Problem?" *Philosophy Compass* 3.4 (2008): 703-720.

Shafer-Landau, Russ. *Moral realism: A defence*. Oxford: Clarendon, 2003.

Shafer-Landau, Russ. "Moral and Theological Realism: The Explanatory Argument." *Journal of Moral Philosophy* 4.3 (2007): 311-329.

Sinclair, Neil. "Recent work in expressivism." *Analysis* 69.1 (2009): 136-147.

Smart, J.J.C. "Sensations and brain processes." *The Philosophical Review* 68.2 (1959): 141-156.

Skarsaune, Knut Olav. "Darwin and moral realism: survival of the iffiest." *Philosophical studies* 152.2 (2011): 229-243.

Stevenson, Charles Leslie. "The emotive meaning of ethical terms." Mind 46.181 (1937): 14-31.

Stiller, James, and Robin Dunbar. "Perspective-taking and memory capacity predict social network size." *Social Networks* 29.1 (2007): 93-104.

Street, Sharon. "A Darwinian dilemma for realist theories of value." *Philosophical Studies* 127.1 (2006): 109-166.

Stroud, Barry. "The charm of naturalism." *Proceedings and Addresses of the American Philosophical Association* 70.2 (1996): 43-55.

Sturgeon, Nicholas L. “Moral explanations.” *Foundations of ethics: an anthology*. Ed. Russ Shafer-Landau and Terence Cuneo. Malden, MA: Blackwell, 2007. 337-352.

Tajfel, Henri. "Social identity and intergroup behaviour." *Social Science Information* 13.2 (1974): 65-93.

Thagard, Paul R. "The best explanation: Criteria for theory choice." *The Journal of Philosophy* 75.2 (1978): 76-92.

Tipler, Frank J. "Cosmological limits on computation." *International Journal of Theoretical Physics* 25.6 (1986): 617-661.

Unger, Ron, and John Moult. "Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications." *Bulletin of Mathematical Biology* 55.6 (1993): 1183-1198.

Unwin, Nicholas. "Norms and negation: A problem for Gibbard's logic." *The Philosophical Quarterly* 51.202 (2001): 60-75.

Unwin, Nicholas. "Quasi-Realism, Negation and the Frege-Geach Problem." *The Philosophical Quarterly* 49.196 (1999): 337-352.

Wheatley, Thalia, and Jonathan Haidt. "Hypnotic disgust makes moral judgments more severe." *Psychological science* 16.10 (2005): 780-784.

Wielenberg, Erik J. "On the Evolutionary Debunking of Morality*." *Ethics* 120.3 (2010): 441-464.

Williams, Bernard. *Ethics and the Limits of Philosophy*. Oxford: Taylor & Francis, 2006.

Wrigley, Colin W. "Giant proteins with flour power." *Nature* 381.6585 (1996): 738-739.

Zoepf, Katherine. “A dishonorable affair.” *The New York Times*. The New York Times, 23 September, 2007. Web. 5 May 2013.