Moral Intuitions in Reflective Equilibrium: Applying Scientific Methodology to Ethics

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF THE
UNIVERSITY OF MINNESOTA
BY

Matthew E. Brophy

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Norman Dahl, Advisor

October, 2009

Dedication

This dissertation is dedicated to Robert J. Brophy, my loving father, whose devoted service as a professor of English at California State University, Long Beach, contributed to my own achievements. I would also like to dedicate this dissertation to my mother, Mary Lou Brophy whose professional background in psychology certainly inspired my philosophical approach. Lastly, I would like to dedicate this dissertation to my daughter, Isabella. Thank you for not arriving before your due date, or I might have missed my dissertation defense. I love you, and look forward to watching you grow up in the many years ahead.

Acknowledgements

With deep gratitude I would like to acknowledge the following individuals. First and foremost, I would like to express my profound appreciation to my advisor, Norman Dahl. His guidance has made me a much better philosopher. Though I would imagine my dissertation is lengthier than average, I suspect that if one were to weigh a compilation of his comments on my various drafts and the dissertation you find here, the scales would likely weigh in his favor. His incisive and extensive comments on my dissertation were invaluable. I would also like to thank Norman Bowie. Though not my official advisor, he was instrumental to my academic and professional development as an ethicist, and taught me that ethics should be applicable to the real world. His comments on my dissertation proposal were illuminating, and his questions during my dissertation defense were stimulating and insightful. I would like to thank the rest of my committee members who participated in my defense: Sarah Holtman and Michelle Mason. During the dissertation defense, they helped me realize that my philosophical approach would receive rigorous objections from more traditional camps; I thank them for their contribution. I would like also to thank C. Kenneth Waters. Before I defected to moral philosophy, he was my favorite professor in the philosophy of science. As you will see in the pages beyond, my dissertation is indebted to my background in the philosophy of science, and much of what is written here stems from my experience in his classes, and my association with the Minnesota Center for the Philosophy of Science. I would also like to acknowledge Paul Tang, Professor Emeritus at California State University, Long Beach; he was a dedicated mentor, and encouraged me to pursue my interests in the philosophy of science in the first place. Lastly, I would like to thank my parents for their unwavering love and support, but most importantly for their faith that, someday, I would finally finish.

Table of Contents

Chapter 1: "Morally Relevant Features and Moral Methodology"

Dissertation Introduction

The interplay between moral judgments and moral principles is a vital part of moral methodology. Most ethical theorists acknowledge this fact, and exhibit deference to moral intuitions: Neo-Kantians are troubled by the Inquiring Murderer; utilitarians by the Framed Innocent Man; virtue ethicists by the Mafioso; and so forth.[1] We seem to recognize that even if we have a map of moral principles, without the compass of moral intuitions, our ability to navigate the moral landscape would be lost. Principles without intuitions are empty, and intuitions without principles are blind.[2]

In this dissertation, I examine the credibility of moral intuitions and their relation to moral principles and background theories, as represented in the method of reflective equilibrium (MRE) originally set out by John Rawls (1971). As part of elucidation and justification of this method, I make frequent comparisons to scientific methodology, which shares close similarities to the method of reflective equilibrium. I argue that MRE provides a non-foundationalist moral methodology, which appears to be a promising approach to moral justification and moral adjudication.

Moral intuitions are a crucial feature of MRE: they serve as the starting points of moral theory construction and testing in a similar way as empirical data serve as the starting points of scientific hypothesis construction and testing. Moral intuitions – just as any data – can sometimes be mistaken, however. Upon what basis can the credibility of a moral intuition be determined? I examine how the credibility of an intuition can be determined by examining its "etiology."[3] The etiology of a moral intuition is its causal

[1] Immanuel Kant countenanced the inquiring murderer example, for instance; nevertheless, our moral intuitions judge that it would be, at least, morally permissible to lie to such an inquiring murderer.

[2] This statement has its source in Kant, who states in the Critique of Pure Reason (1998, B 76), "Thoughts without content are empty, intuitions without concepts are blind." Another quote that is also often attributed to him states: "Experience without theory is blind, but theory without experience is mere intellectual play."

[3] Richard Joyce (2006) deems the causal origins behind an intuition a "genealogy," concerning where the intuition came from.

origin, which includes sociological, psychological, evolutionary and biological factors, some of which might impugn its credibility.

Since intuition credibility determination is essential to the methodology of reflective equilibrium, I endeavor to show that moral intuitions can be vetted in nontrivial and noncircular ways. This filtration process discredits those initial moral judgments that are determined to be error-disposed. These resulting noncredible intuitions are then excluded from the set of considered judgments, which serve as the provisional starting points for ethical theory construction and testing.[4]

We commonly acknowledge error-disposed conditions as we morally navigate our ordinary lives. If we have a moral judgment, and suspect that some error-disposed conditions are present, we often refrain from moral deliberation or seek to "massage" our intuitions, and then engage in deliberation a different way.[5] One common error-disposed condition we recognize is emotionality.[6] In some cases, emotionality seems to influence unduly our moral judgments. Empirical research in experimental moral psychology can help determine the occasion and degree of the influence various emotions have upon our judgment. In addition to psychological factors, biological, evolutionary or social factors can also affect moral intuition credibility by unduly influencing our moral perceptions of the world.

Ultimately, I will show that the moral methodology of reflective equilibrium, when theoretically developed and empirically substantiated, provides a significant contribution to moral philosophy. In particular, this fortified methodology provides further traction in ethical debate and adjudication. I exemplify this point in the final chapter, demonstrating how intuition credibility determination lends defense to a certain form of utilitarianism against certain traditional intuition-based attacks, and I show how the triple adjustment between intuitions, moral principles and background theories, understood in the context of wide reflective equilibrium, can assuage such objections.

---

[4] I am using "intuitions" and "judgments" synonymously. I employ the term "intuitions" when speaking of moral judgments generally, without distinguishing between initial judgments and considered judgments.

[5] One way to "massage" one's intuitions is to consider the adage, "What if I were in another person's shoes?" This is supposed to help one overcome bias or spur sympathy by providing a different perspective than might be myopic in its perception.

[6] Emotionality can also be a credibility-amplifying condition, something I will discuss later.

For the purpose of providing an overview of this dissertation, I will provide a chapter-by-chapter roadmap of it. Before I embark on these synopses, I believe a brief snapshot of the project can be provided by referencing the beginning paragraph of Gilbert Harman's article, "Three Good Trends in Moral Philosophy." This passage serendipitously sums up three primary aspects of my dissertation. For the purposes of elucidation (and certainly not as an appeal to authority, I'm sure), I present Harman's introductory paragraph, in full:

> "There have been three good trends in moral and political philosophy over the last fifty years or so. First, there has been a trend toward rejecting special foundations, a trend that is exemplified by the widespread adoption of the method John Rawls adopts, in which particular judgments and principles are adjusted to each other in an attempt to reach 'reflective equilibrium.' Second, there have been attempts to use intuitions about particular cases in order to arrive at new and often arcane moral principles like that of double effect, as in discussions of so-called trolley problems. Third, and perhaps most important, there has been increased interaction between scientific and philosophical studies of morality, as for example in philosophical reactions to psychological accounts of moral development and evolutionary explanations of aspects of morality" (2003, p. 415).

I discuss the first trend, methodologies of reflective equilibrium, in chapter two; the second trend, intuitions abstracting to moral principles, I explore in chapters two, three and five; and the third trend – the interaction between scientific and philosophical studies of morality – I examine in chapter four and throughout the project. In order to contribute to an overall understanding of the dissertation, the following subsections provide a brief overview of each chapter.

Chapter Overview

In chapter 1, I will provide some basic concepts and theses that will prove useful for the rest of the dissertation. In particular, I will characterize my understanding of the concept of moral intuitions; as part of this, I will set out the social-intuitionist model of experimental moral psychology, and delineate the way in which a moral intuition can be considered morally relevant or irrelevant. In chapter 2, I present and explain the method of wide reflective equilibrium, ultimately showing why wide reflective equilibrium is superior to the narrow method in that it has the additional resource of background

theories in reflective adjustment. As part of the method of reflective equilibrium, generally, I focus on the filtration process, showing how this process can diminish the credibility of certain moral intuitions, if the etiology causally responsible for the generation of the intuition is found to be suspect. In chapter four, I provide a variety of empirical research across several scientific and social-scientific fields, which bear upon intuition credibility: I argue that certain etiologies do provide significant reason to diminish the credibility of certain intuitions. While chapter three argues that credibility can *in principle* be diminished due to certain etiologies, chapter four argues that in certain cases credibility *actually* will be diminished due to various etiologies. In chapter five, I put to work this entire moral methodology of wide reflective equilibrium. Specifically, I show how this methodology can be used to defend utilitarianism against certain traditional objections, by employing the filtration process, as well as reflective adjustment via background theories, and by making adjustments to the theory itself. This chapter will not, of course, provide a complete defense of the form of utilitarianism I promote in this chapter, but it will provide a significant part of such a defense, and it will illustrate how a fortified method of wide reflective equilibrium can advance the adjudication of moral theories.

Some Basic Concepts and Theses

As I have already indicated, in the remainder of this chapter I introduce certain concepts and arguments fundamental to the dissertation. First, I explore the nature of moral intuitions, and the functioning of moral cognition moral generally. Currently, empirical research in cognitive science appears to be supplanting the traditional "rationalist" model of moral cognition with what is called the social intuitionist model. I describe this model, and indicate what bearing it has upon the understanding of our moral judgment. Second, I attempt to substantiate the distinction between morally relevant and irrelevant features, which is an essential distinction employed by the method of reflective equilibrium. In one argument for this distinction, I proffer that there are certain ways

human beings must understand morality, which I term "ethical forms of life."[7] As part of the argument for this distinction, I appeal to shifting the burden of proof, as well as the Rawlsian concept of overlapping consensus. Last, I briefly consider two objections to this account.

What are Moral Intuitions?

Everyone has moral intuitions. When people refer to gut-feelings, visceral reactions, or twinges of conscience, they are typically referring to the deliverances of their moral sense.[8] "Moral intuition" denotes this putatively pretheoretical moral judgment that we have about certain moral objects: murder, rape, cruelty, helping others, and other moral matters.

Historically, moral intuitions have been often likened to perceptions. Moral intuitions were conceived to be immediate "observations" of moral truths in the world.[9] Intuitions have also been compared to aesthetic judgments, where one has an affective reaction toward an object: for instance, viewing a painting of dogs player poker might cause one to have an immediate negative feeling related to this art object. This latter interpretation, where moral intuitions arise much like aesthetic judgments in being valenced evaluations, has been substantiated by current empirical research in experimental moral psychology, and comports with the characterization of moral sentiments by David Hume (1751, p. 13): moral judgments are arrived at via "immediate feeling and finer internal sense" rather than by "chain of argument and induction."

Both analogies – intuition as perception and intuition as aesthetic judgment – share the same primary features: intuitions are relatively immediate, tend to be automatic, and are not consciously derived. To briefly contrast moral intuitions with moral

[7] I discovered that this term is similar to the term "moral forms of life," that has been used by Margaret Walker (1998); she uses it in a similar way in her work in feminist ethics and epistemology. I will define this term more thoroughly later in the chapter.

[8] I don't wish to deem one's "conscience" as synonymous with moral intuitions, but I would suggest they are interrelated. I won't explore the concept of conscience in this project, however.

[9] Though intuitions are typically immediate, at times we need to observe or consider the object for a while before an intuition arises, just as we might have to observe/consider an object in the visual world for a while before we can make out what it is.

derivations, consider the following example. Imagine Sue happens upon Peter, who is kicking a dog. Sue is immediately horrified, in a moral sense, in that she judges that the action Peter is engaging in to be immoral. This isn't a determination that she arrives at via a derivation from moral principles, but is a judgment that immediately arises within her. Alternatively, it could be the case that, instead of intuiting this action as wrong, Sue deduces Peter's action to be wrong from moral principles (for instance, that it violates the Golden Rule).[10]

The contrast between moral intuitions and moral reasoning, then, seems to be that an intuition is an immediate moral determination that occurs w/o conscious deliberation. Though moral intuitions might be subsequently presumed to be "pretheoretical," this would be a mistake. Moral intuitions arise under some theoretical assumptions, though these moral judgments are still not theoretically derived on a conscious level. When the moral judgment arises in Sue that Peter's kicking a dog is immoral, it may be based upon her conceptual understanding that dogs are sentient beings, for instance. In such a way, her moral intuition is not "theory-free."[11]

That moral intuitions are theory-laden should not be considered an insurmountable problem for a moral methodology that incorporates them. The same is true of observations in science; all scientific observations necessarily presume some theory. Even ordinary sensory perception presumes some theoretical presumptions: for instance, when a congenitally blind person gains sight later in life, he cannot immediately make sense of the visual world he apprehends without first learning an appropriate way to interpret it.

---

[10]In analogy to science, an astronomer might simply "observe" via telescope where a moon is its orbit around a planet, or the astronomer can deduce where a moon is through mathematical calculations of gravitation and positions and velocities of other objects in the same system.

[11] It might be the case that the individual has internalized a moral principle, where the moral judgment is a manifestation of this principle. For instance, "Do not hit others" might be a principle via moral education that results in the moral judgment that Peter's kicking of the dog is immoral. In this case, we would have to investigate the MRFs of this "subconscious" principle. It needn't be problematic that moral judgments can be instances of internalized moral principles, as they may vary in complexity. In analogy with science, data can vastly differ in its dependence upon theoretical assumptions. Consider, for instance, the theoretical assumptions in searching for a planet in our solar system using a telescope, or even through calculations via Newtonian mechanics, versus searching for a quark, which depends on more assumptions.

Current research in moral cognition has distilled the concept of moral intuitions further. Jonathan Haidt makes the distinction between moral intuitions and moral judgments (2001, pp. 814-820); I, however, will be employing the two terms synonymously. To avoid confusion, I will refer to Haidt's concept of moral intuitions as "proto-intuitions." I will limit the discussion of proto-intuitions to this chapter.

Proto-intuitions, according to Haidt, are pre-conscious, affectively valenced evaluations. An empirical illustration of proto-intuitions are the valenced evaluations we generate when assessing race, as demonstrated by the famous Harvard Implicit Association Test.[12] The most popular test in this series of tests pertains to associations with race, where the subject is directed to categorize certain words and faces with one of four categories: African American and Good; African American and Bad; European American and Good; and European and Bad. The subject's performance is tracked for both time and accuracy. Typically, test subjects (of all races) have a far easier time associating African American with Bad and European American with Good, than vice versa. This study seems to suggest that we harbor subconscious affectively valenced reactions toward different races. Though we might not consider ourselves racially biased, it appears that at some level we just may be. However, this proto-intuition is not the end of the story: there are also other processes that may come into play that affect the output resulting in a conscious judgment. For instance, we could also have certain "correctives" in our cognition, where other proto-intuitions can combat the initial negative association. The end product of a conscious moral judgment may be all that we are aware of, and yet there may be several steps, challenges, and revisions it took to get there.

Proto-intuitions, as affectively valenced evaluations or attitudes, could have significant influence upon resultant moral judgments; for this reason, we might seriously consider how best to alter these proto-intuitions.[13] For instance, the above evidence

[12] This test can be taken on the Internet at https://implicit.harvard.edu/implicit/demo/index.jsp

[13] Extended research of the Harvard Implicit Association Test showed that subjects who spent an hour, prior to taking the test, studying Black history, which focused upon positive role models, had a much easier time making positive associations with African American faces than otherwise (Gladwell 2005, p. 97). I would speculate that, similarly, portrayals in the cinema, television and music of African Americans that associate them with negative qualities have an influence on us, though we might not be consciously aware of these effects.

regarding implicit association should give us some pause, as presumably our proto-intuitions – if affectively valenced in such a way – may lead us toward actual racial bias in attitude and action. For instance, if African Americans were more easily associable, in one's mind, with negative concepts, this might have some bearing upon how we assess the group morally: such as, the evaluation that African Americans deserve less charitable assistance as a group than do European Americans.[14]

Further empirical evidence substantiates the undue influence of certain attitudes upon moral judgments. For instance, a psychological phenomenon called the "halo effect" describes where positive evaluations of nonmoral traits unduly influence subsequent moral evaluations. One 1972 study (Dion, Berscheid, & Hatfield) found that an aesthetic evaluation by a subject that a person is attractive leads to beliefs that that person also has "corresponding" positive moral traits, such as kindness and good moral character. Presumably, however, attractiveness is not a morally relevant feature upon which to base such moral character assessments, as it is false that attractive persons possess more positive moral traits than those who are judged not to be particularly attractive.

To further elucidate moral cognition, and our faculties of moral judgment, I introduce and examine in the next section the social intuitionist model, as presented by Jonathan Haidt (2001). This model is substantiated by recent cognitive studies of moral reasoning, and suggests that emotions may be prior to rationality in the generation of moral intuitions. The significance of the social intuitionist model is its reordering of our traditional understanding of how human beings cognitively generate moral judgments, where emotionality is primary, and rationality is secondary. In addition, the social-intuitionist model exposes how emotionality can dominate our rationality when we generate moral intuitions; this may, at times, significantly dispose our moral judgment to error. The importance of examining moral cognition is to help identify and avoid hazards in moral judgment and coherence adjustment in MWRE.

[14] Additional studies seem to indicate a "hostile" reaction toward African Americans, whether by associated words or subliminally presented pictures: Devine, 1989; Bargh, 1996; respectively.

The Social Intuitionist Model

Current empirical research in moral cognition is leading to an overturning of traditional rationalist models concerning how people reach moral judgments. The social-intuitionist model, presented by Jonathan Haidt, suggests that moral emotions heavily influence valenced proto-intuitions, which lead to moral judgments, after which post-hoc rationalizations are searched for in justification of the moral judgment (Haidt, Bjorklund, & Murphy, 2004). This is an inverse of the rationalist model, where rationality supposedly determines moral judgment, at which point an individual's moral feelings follow in reaction to that decision.

In illustration, consider moral judgments about homosexuality. According to the Social-Intuitionist Model, an individual would generate a moral judgment in the following way: First, he would behold an eliciting situation, such as viewing a picture depicting two men kissing; second, he would experience a moral emotion, such as disgust; third, he would subsequently judge the prospect as immoral; fourth, he would then seek to justify that moral judgment with rational reasons: that homosexuality threatens to lead to bad consequences for society; that it's "unnatural" and therefore immoral; and so forth.

James Rachels (1986, ch.1) considers the reasons people give for their moral condemnation of homosexuality, and determines them to be unjustified, illogical, or specious. Nevertheless, despite cogent counterarguments to such justifications, it seems almost impossible to dislodge someone from their condemnation of homosexuality, even in the face of compelling arguments. This goes to support further the social-intuitionist model of moral cognition: emotions determine moral judgment, and those emotions are in large part determined by our social interactions with others, and what they approve of disapprove of morally. Haidt characterizes our moral reasoning, via simile, as more akin

to a lawyer seeking to justify their case than an impartial judge, dispassionately weighing all the evidence before making a determination (Haidt, 2001, p. 814).

That our emotional faculties can often unduly dominate our rational assessments is empirically evident in political judgments. Consider a recent study exposing confirmation bias in individuals' political evaluations (Westen, Kilts, Blagov, Harenski, & Hamann, 2006). In the study, the brains of staunch Democrats and Republicans were monitored when asked to assess information contrary to their ideology. The emotional centers of the brains were much more active than the reasoning centers: subjects would experience negative emotions when confronted by evidence contrary to their political ideology, followed by emotional reward when they resolved this ostensible contradiction by seizing upon information that might explain away the contradiction. Westen, author of the study, reports that subjects would reach significantly biased conclusions by ignoring information contrary to their point of view that could not be rationally discounted, while remaining cognizant of evidence contradictory to the viewpoints of the opposing political party. That is, both Republicans and Democrats consistently denied obvious contradictions for their own candidate while detecting contradictions in the opposing candidate. Westen concludes: "The result is that partisan beliefs are calcified, and the person can learn very little from new data" (Clark, 2006, p. 1). This study would seem to suggest that one's emotional disposition toward another individual, in this case a candidate, affects one's normative evaluation of that individual in a way which is error-disposed. Other related studies suggest that we're not only biased when it comes to evaluations of others, but also of our selves: that is, while we are often able to assess bias in others accurately, we tend to be self-deceived to the fact we are often, ourselves, biased (Ditto & Lopez, 1992; Miller & Ratner, 1998).

The default presumption of how our moral judgment functions is that disagreeing individuals have different factual starting assumptions which consequentially lead to different moral judgments. A study, however, by Haidt, Koller, and Dias (1993), and a second study by Haidt and Hersh (2001), both found evidence that suggests the actual causation is the reverse of this: for example, that the moral judgment that anti-abortion is

unethical causes the factual belief that life begins at conception, rather than vice versa.[15] Surprisingly, affective reactions were much better predictors of moral judgment, overall, than assessments of potential harmfulness.

Such evidence supports the social-intuitionist model that valenced proto-intuitions lead to moral judgments that are followed by post-hoc reasoning. If the social-intuitionist approach is right, then this leads to the worry that emotionality may more readily lead to bias in our moral judgments and reasoning. This supports Rawls' claim that emotionality is a condition that needs to be considered in assessing the credibility of moral judgments.

In investigation of emotion and moral judgment, cognitive neuroscientists have begun to study the correlation between humans' brain activity when considering moral dilemmas. In a 2001 article, a group of five neuroscientists conducted two functional magnetic resonance imagining (fMRI) studies, utilizing moral dilemmas as "probes," to study moral judgment via the methods of cognitive neuroscience. Their abstract states:

> "We argue that moral dilemmas vary systematically in the extent to which they engage emotional processing and that these variations in emotional engagement influence moral judgment. These results may shed light on some puzzling patterns in moral judgment observed by contemporary philosophers" (Green, J., Sommerville, R., et. al. p. 2105).

In support of the social intuitionist model, this empirical evidence via fMRI experiments indicates that our moral intuitions are primarily influenced by the emotional centers of our brains. I will consider some of this empirical evidence in the following section.

Experimental Moral Psychology: Testing the Social-Intuitionist Model

In a study of moral cognition, led by psychologist Joshua Greene, subjects were asked to make moral judgments concerning hypothetical trolley car cases, while their brain activity was monitored via fMRI (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001).

[15] These studies are, in effect, responses to a previous 1991 study by Turiel, Hildebrandt, and Wainryb that examined young adults reasoning about abortion, pornography, homosexuality, and incest.

The moral dilemma the neuroscientists presented to their subjects was a series of trolley examples.[16] Typically, when subjects are presented these examples – whether they are experimental subjects in professional studies or undergraduates in introduction to philosophy courses – they strongly tend to endorse pulling the lever in the trolley case, which diverts the runaway trolley car onto an alternate track, which would kill one to save five, and yet condemn pushing a person off of a footbridge to stop the runaway trolley car, which would also kill one to save five. Yet in both cases, the morally relevant features appear to be the same: the subject kills one person to save five people:

> "This leaves psychologists with a puzzle of their own: How is it that nearly everyone manages to conclude that it is acceptable to sacrifice one life for five in the trolley dilemma but not in the footbridge dilemma, in spite of the fact that a satisfying justification for distinguishing between these two cases is remarkably difficult to find?" (Greene, et al., 2001, p. 2106)

The social-intuitionist model, as an empirical model delineating moral cognition, does not make any claims in regard to the presence or absence of morally relevant features. Nevertheless, the disparity in moral judgments between the two ostensibly identical cases does present a challenge to ethical approaches that rely upon the credibility of moral intuitions. If the moral judgment that arises in the switch case is different than the moral judgment that arises in the footbridge case, then this discrepancy needs to be justified via a difference in morally relevant features, or in some error in one or both cases. I will explore this latter possibility in chapter five: I will argue that the two cases are not actually identical in their morally relevant features.

Greene and colleagues conclude that the crucial difference in moral cognition between the two trolley cases is that the footbridge dilemma (when subjects are given the option of pushing the person) engages people's emotions in a way that the trolley dilemma, involving merely flipping a switch, does not (Greene, et al. 2001, p. 2106). They conclude that the far greater activity in the emotional centers of the brain in the footbridge case, compared to the switch-flip case, affects people's moral judgments.

---

[16] Typically, the trolley example presents individuals with two different cases: one where you flip a switch to kill a person to save five people; the other where you push a person to his death in order to save five people. Test subjects morally intuit that flipping the switch to save the five is morally permissible whereas pushing the person to his death to save the five is impermissible. The morally relevant features of in the two cases ostensibly appear to be identical, however. I will discuss in chapter five whether or not the two cases are in fact identical in terms of their MRFs.

Their data suggest that judgments concerning "impersonal" moral dilemmas more closely resemble judgments concerning nonmoral dilemmas than they do judgments relating to "personal" moral dilemmas (Greene, et al., 2001, p. 2107).

From the social-intuitionist model, we learn that emotions can have a significant influence upon our moral judgments. This doesn't necessarily indicate that our moral intuitions, when influenced by emotions, are not credible; nonetheless, it does necessitate a greater focus upon emotions in moral cognition. In the past decade, there has been an increase in academic examinations of the morality-related emotions – shame, guilt, contempt, regret, disgust: particularly, there has been a resurgence of virtue ethics in the field of moral philosophy. This field, in part, inquires what conditions need be necessary for an emotion to be appropriate or justified.[17]

Since emotions have such an influence on our moral judgments, rather than (*post hoc)* moral reasoning which appears to be secondary, we need to consider emotions as a primary area of interest: investigating when they're appropriate and when they are not appropriate. Philippa Foot, for instance, presents the example of a person being, inappropriately, proud of the sky: pride, as an emotion, is only possible given certain necessary conditions. In a related and more general way, I want to argue that in order for a moral judgment to be deemed credible, it needs to be based upon morally relevant, rather than irrelevant, features. If a moral judgment is based upon, or significantly influenced by, certain emotional conditions – such as when an influencing emotion is present during the generation of a moral intuition, but unrelated to the object being morally evaluated – then that judgment should be attributed diminished credibility.

The emotional conditions I'm referring to count as what I term "error-disposed conditions." Rawls recognizes emotionality as an error-disposed condition:

> "We can discard those judgments made with hesitation, or in which we have little confidence. Similarly, those given when we are upset or frightened, or when we stand to gain or lose one way or the other can be left aside. All these judgments are likely to be erroneous or to be influenced by an excessive attention to our own interests." (1971, p. 47)

[17] For an example, see "Contempt as a Moral Attitude," Mason, 2003.

Emotions can both amplify or diminish credibility. Emotionality diminishes the credibility of an intuition if it causes the intuition to be based on morally irrelevant features. For a moral intuition to be credible, it needs to be based upon morally relevant features. Error-disposed conditions include emotionality, but also extend to other sources of possible error. I examine these other sources of error in chapter three; I also consider certain conditions that should be taken as amplifying credibility. To substantiate the moral relevance distinction, which undergirds error-disposed conditions, I present varied argumentation in the following several sections.

Determining Moral Relevance

One primary focus of this dissertation project is to explicate reasons by which to determine whether an intuition is sufficiently credible or is noncredible: namely, that an intuition is credible to the degree it is based upon morally relevant features (MRFs) rather than morally irrelevant features (MIFs). Subsequently, the question naturally arises: What makes a feature – in regard to an action, person, object, or situation – morally relevant?

This is a weighty question, for which I will attempt to provide a few different but interrelated answers. My hope is that one or more of them, singularly or in tandem, will prove sufficient for the purposes of this dissertation. In attempting to answer this fundamental question, I will consider the following topics in defense of my use of this distinction:

1. Commonsense morality and the extreme poles of morally relevant features and morally irrelevant features.

2. Morally relevant features as inherent and inexorable to our perception, grammar, and conceptual understanding (i.e., ethical forms of life).

3. Shifting the burden of proof onto those who would deny this distinction to justify their denial while remaining moral subscribers who subscribe to moral objectivity

4. An overlapping consensus, similar to Rawls' initial intuitions of justice.

Commonsense Morality and Moral Relevance

Before delving in search of philosophical justification of the moral relevance distinction, it may provide some *prima facie* support to note that we already have a relatively clear commonsense understanding of at least part of what constitutes moral relevance and irrelevance. This is most apparent at the extreme poles.

In illustration of these poles, let's consider the simple example of kicking a dog. Imagine that Phillip doesn't kick innocent dogs because he recognizes that dogs are sentient: they can feel pain. Imagine that Kyle, in contrast, doesn't kick innocent dogs because he recognizes that dogs are furry. In our ordinary world, it's clear that the fact dogs are sentient is a morally relevant feature, whereas the fact dogs are furry is a morally *irrelevant* feature. Both intuitions may happen to be correct – kicking innocent dogs is immoral – but only the intuition based upon the MRF of sentience is credible.[18]

In a similar way, briefly examine the case of slavery: surely, the suffering of slaves is a morally relevant feature in an individual's moral judgment that slavery is intrinsically wrong. Contrarily, if it happens that slavery is financially beneficial to a given individual's peers but not to his own wealth, this is not a morally relevant feature for that individual to judge that slavery is intrinsically wrong. The reason for this difference is that suffering is something intrinsic to the wrongness of slavery, whereas an individual's being left out of financial benefit is not something intrinsic to the wrongness of slavery, but is rather an extrinsic and contingent feature.[19]

These two examples underscore that we do recognize certain features as clearly morally relevant and others as clearly morally irrelevant.[20] The relevancy determination may be more difficult to make in other intermediary cases, where there could be disagreement.[21] I will not consider such controversial cases in this project, though

[18] There might be other MRFs in addition to sentience, such as, possibly, intelligence. Peter Singer does not regard intelligence as an MRF, but his arguments against this feature as morally relevant seem specious. However, I will not argue against his dismissal, here. Suffice to say: There can be controversy about what features are morally relevant or irrelevant without eroding the broader distinction.

[19] Though the individual might be able to intelligibly claim that being bereft might be an extrinsic property of slavery which makes it wrong in terms of unfairness: peers are benefiting but you are not.

[20] I believe such understanding extends internationally, cross-culturally, and historically.

[21] For example "purity" versus "corruption" as intrinsically morally relevant, rather than just as a

determining what features are or are not morally relevant in such cases is certainly an important research program, worthy of pursuit. For the purposes of my argument concerning intuition credibility determination, it should suffice to note that we do understand, recognize, and accept that there are at least some clear cases of moral relevance and moral irrelevance. This distinction can provide a basis upon which to assess the credibility of moral intuitions.

Even in these polarized cases, it might be pointed out that oftentimes immediate clarity is lacking. A white supremacist may feel that the color of an individual's skin is morally relevant in regard to how that individual must be treated. Though skin color might be what immediately sparks the supremacist's intuition, the intuition is actually based upon other alleged features. If we pressed the white supremacist, he would likely acknowledge that skin color is not the intrinsic feature upon which he is basing the moral inferiority of non-whites. Skin color is just a marker, which, he might assert, denotes morally relevant features: such as his belief that non-whites possess an animalistic nature, a lesser intelligence, a supposed unholiness, impurity or criminality, or other such poppycock.[22] We might clarify the white supremacist's position by entering interlocution with him, and presenting him with certain counterexamples. For instance, we might have him imagine that the skin color of one of his white supremacist's compatriots changed to dark brown whereas the compatriot's history and faculties remained the same, and then ask whether or not this would be sufficient to make his compatriot inferior: presumably, he would find this superficial alteration insufficient to reduce the intrinsic moral considerability of his compatriot.

Counterexamples are one helpful tool by which to clarify morally relevant features from morally irrelevant features. Peter Singer (1975) employs numerous counterexamples in his work concerning the ethical treatment of animals to illuminate what features, which we might initially cite as morally relevant, are in fact morally irrelevant. In many instances, his counterexamples are compelling and convincing, which

---

marker for some other MRFs.

[22] These, in turn, are terms related and, perhaps arguably, reducible to intelligence, sentience, or some other elemental concept.

is one of the reasons why his book, *Animal Liberation*, has been so widely influential and revelatory to many.

In his book, Singer trots out candidates who are oftentimes taken to be morally relevant features. For instance, in defense of our moral anthropocentrism, we might cite having human DNA as morally relevant for intrinsic moral considerability. However, if this is in fact a relevant basis, we would have then to grant equal intrinsic moral considerability to severely and permanently mentally vegetative humans, while denying it to higher mammals (e.g., dolphins and whales) as well as denying it to hypothetical alien beings who exhibit advanced mental and moral capabilities without having a scratch of human DNA.[23]

Via such arguments from absurdity, we can expose when the features upon which a moral judgment is based are morally relevant or irrelevant. If we determine those features turn out to be irrelevant, we can either scramble to find other morally relevant features in justification of our moral judgments, or we can alter those moral judgments in deference to those features which remain.[24]

When such features, such as skin color or villosity, are purported to be the basis for moral judgments, it needs to be examined further if they are in fact serving as markers, or if a "special background" is being presumed.[25] In this special background is where the morally relevant features actually reside.[26] I examine the moral relevance distinction further, below, and attempt to establish a philosophical argument that this

[23] In his interrogation of our moral assumptions, Singer seems to be challenging people's personal WRE sets. Some cite the ability to use language as a morally relevant feature: however, if this were the morally relevant feature to predicate upon, Singer argues, we would have to include chimpanzees, dolphins, and possibly future artificial intelligence, but exclude infants, some elderly, and the severely mentally retarded.

[24] According to Singer (1975), the only remaining MRF seems to be sentience. Acceptance of sentience as the only MRF would demand a change, on pain of irrationality, of our evaluation of the moral considerability of animals as well as humans.

[25] The notion of "special background" is presented by Philippa Foot (1959), which I will expound upon later in this section.

[26] For instance, a royal heir might claim right to rule over others based on their name, but this is not a sufficient MRF; in actuality, the "special background" proffered to morally justify such rule and superiority is, for example, that the name/heritage is a marker denoting divinity, intrinsically possessed or extrinsically bequeathed as serving as a legitimate agent of the divine. If not based expressly on divinity, it is based upon some notion of superiority: e.g., intrinsic superiority of intelligence and sophistication to those of the unwashed and base masses.

distinction is justified in that it is an inexorable and necessary part of human moral understanding.

Morally Relevant Features and Ethical Forms of Life

There are two questions that can be raised in challenge to the concept of morally relevant features: Can't *any* feature be deemed morally relevant? And, inversely, can't any feature we commonly deem to be morally relevant just as easily be deemed as morally irrelevant by someone else? I believe the answer is no to both questions. I intend to show that there are certain "ethical forms of life," which are necessary to morally understand the world, if we are to understand the world in regard to morality at all.

"Forms of life" is a Wittgensteinian concept relating to language: it is what enables language to function, and therefore must be accepted as a "given." My assertion is that in order for us to be moral subscribers at all, we must recognize certain ethical forms of life while rejecting others. Though Wittgenstein didn't flesh-out the idea much, himself, in reference to language, I hope the concept I wish to employ will become clear through the discussion in the next three sections. I think that the concept of ethical forms of life can be inferred from Foot's work concerning moral conflict and the internal relation of nonmoral concepts with moral concepts.

One note regarding terminology in the discussion that follows: I will be using the term "moral subscribers" to denote those who subscribe to morality in an ordinary sense. Moral subscribers take moral propositions to be consistently and objectively correct or incorrect. For instance, the proposition "torturing children just for fun is wrong" is correct if the act of torturing children just for fun is, in fact, wrong. In this way, moral subscribers are not moral skeptics. My assertion is that any moral subscriber must operate in accordance to ethical forms of life. If he does not, then whatever he is doing isn't recognizable as morality – not by fiat or majority rule, but by the fact that it is unintelligible to anyone else (and perhaps even to himself) as morality.[27]

[27] I will elucidate this claim further at the end of the chapter via the example of the happy torturer.

The Challenge of the Moral Eccentric

Philippa Foot considers two questions closely similar to the ones posed at the start of this section. She entertains the contention:

> "…a moral eccentric could argue to moral conclusions from quite idiosyncratic premises; he could say, for instance that a man was a good man because he clasped and unclasped his hands and never turned N.N.E. after turning S.S.W. He could also reject someone else's evaluation simply by denying that his evidence was evidence at all" (1959, p. 111).

Relating this point to our discussion, the contention would be that the moral eccentric could argue to moral conclusions from what we deem "morally irrelevant" features – such as clasping his hands and refraining from engaging in certain turning sequences of his body. Moreover, he could reject another person's insistence that "suffering" was a morally relevant feature; in fact, he might maintain that only hand-clasping and body-turning were morally relevant.

Foot unpacks two assumptions about evaluations regarding moral relevance:

(1) "Some individual (or culture) may, without logical error, base his beliefs about matters of value entirely on premises which no one else would recognize as giving evidence at all" (p. 84)
(2) "…given the kind of statement which other people regard as evidence for an evaluative conclusion, he may refuse to draw the conclusion because *this* does not count as evidence for *him*" (p. 84). [28]

I will separately consider both of these assumptions in the five sections, following.

Can One Correctly Base Moral Judgments on Anything at All?

Foot asserts that the first assumption, above, is incorrect because there are constraints upon what an individual can base his beliefs: there is an *internal* relation between beliefs and the premises upon which they predicate (1959, p. 111).

[28] *Her emphasis.*

In illustration, Foot entertains the emotive concept of pride.[29] There are constraints regarding of what a person can sensibly be proud. For instance, consider an individual who claims to be proud of the sky. He might, in fact, feel a sensation of pride. Pride, however, isn't merely a sensation or emotion, it necessitates at least partial accompaniment by beliefs or assumptions of a certain kind. Can an individual be proud of the sky? In order for this to be conceivable to us, we must imagine the man is under some kind of delusion: he saved the sky from falling; perhaps he prevented its becoming polluted; he may believe he created the sky just this morning. In such a case, we can imagine the man is indeed proud of the sky; nonetheless, he is mistaken about his facts, in which case we can criticize him for holding false beliefs. In such a case, we could imagine his pride as not actually being related to the actual object, but a confusion of what the object is: mistaking the preexistent sky for the sky as created by him.

A similar example we might consider is that of a cuckolded father who feels pride about his daughter having his striking green eyes, when in fact, unbeknownst to him, the daughter's green eyes actually have been genetically bestowed by the mailman. We can accept the cuckolded non-biological father as being proud of his daughter's eyes, but claim his pride is misplaced, as the object his pride refers to is significantly different than what he presumes, and is not, it turns out in this case, an appropriate object to be proud of (quite the opposite, perhaps).

Foot remarks that it is striking that, in order to make sense of the man being proud of the sky, we have to introduce background assumptions or stories, which she terms "special background." It is inconceivable that a man could be proud of the sky, full stop. Rather, a special background is a necessary condition, and must be inserted, for pride to be possible (not justified, just comprehensible). This is because pride is a concept that has an internal relation to its object: for example, that you have contributed to the object of pride in some way; also, that the object is positive in some way to someone (e.g., you can't be proud of something you don't endorse on some level and to some degree).

[29] Foot utilizes other examples as well: fear requires appropriate emotion, as well as belief of the object as dangerous; dangerous necessitates the risk of injury or harm of some sort, and serves as a warning function; injury requires some loss of normative function.

In a similar way, some features in moral valuation are not morally relevant features: we cannot predicate moral valuation upon any features, willy-nilly: that is, to restate assumption 1, “an individual cannot base his beliefs about matters of value entirely on premises which no one else would recognize as giving evidence at all” (Foot 1959, p. 111)

Consider the example of an individual trying to justify his rule over other individuals: invading their homes and seizing their property, for instance. The justifications for this behavior are finite in category: typically they involve divine authority of some kind, alleged superiority in faculties, etc. The justifications proceed from assertions of a few special backgrounds. No rulers allege justification to rule over others on the basis of simply being shorter or stockier than his subjects, unless of course such characteristics are a marker for an aforementioned special background.

Eye color, to consider another example, is not a feature someone can morally valuate upon. Imagine that an individual morally valuates Gabe as morally superior to Bruce just because Gabe has green eyes whereas Bruce has brown eyes. Or, if in action, Gabe and Bruce perform the same immoral act, and this individual judges Bruce more morally culpable than Gabe, due to the difference in eye color. Eye color, it strikes us, is just not a relevant object upon which to predicate certain moral features. In order to make this difference in moral valuation stick, we need to introduce a special background: for example, perhaps green-eyed people are believed to in some way superior in faculties than brown-eyed people.

Moral evaluation involves feeling and belief, as illustrated in Foot’s example of an individual asserting that he is proud of the sky. Her example regards an attitude, but is supposed to be analogous to making a moral judgment. One cannot sensibly make any moral judgment about anything he wants; there are certain constraints already in place. For instance, consider the attitude of moral condemnation at an object which we evaluate to be morally blameworthy. With this condemnatory evaluation there are internal relations to the object already: for instance, in some cases that the object of our condemnation is willfully responsible for a violation or injury of some kind. The object of our condemnation, then, cannot be a rock; we cannot find it morally culpable – even if

it impacts one's body and causes injury. We hold the rock-thrower morally responsible, not the rock itself.

Persons may be limited in culpability according to features they possess: for instance, a small child who throws a rock we may hold less morally culpable than a full-grown adult in possession of all his faculties. The child may not have realized the consequences, may not have been morally trained, may lack rational and moral faculties, and may not fully grasp the self/other distinction. These features are, then, morally relevant features in that they appropriately impact our moral judgments of the object.

Moral blameworthiness, and accompanying moral indignation or condemnation to the object that is blameworthy, already presumes choice or the capacity of choice. Glancing at the determinism controversy, we see that compatiblists expend much of their time arguing that there is choice in a sufficiently robust sense. They don't throw up their hands, deny free will, and say "so what," conceding humans have no choice, while still maintaining the assertion that humans are morally culpable. Such a position is one no one holds, as it is an incomprehensible position. For instance, we can't imagine someone holding as morally blameworthy the hail that falls from the sky. If someone *were* to hold this position, we can imagine a special background would be present: such as that hail has intents, beliefs, motives, desires, consciousness, etc. In summary, moral blameworthiness pre-necessitates the capacity for choice as its internal object. Without choice, of some robust kind, we cannot conceive of an object being morally blameworthy.[30]

Moral considerability has necessary internal relations to objects as well. For instance, for an object to be intrinsically morally considerable we presume the object must be conscious on some level (or have potential to be so, such as a fertilized egg or temporarily comatose person). Many objects may be *extrinsically* morally considerable,

[30] In counterpoint, one might reflect on the Christian concept of original sin. Are we morally blameworthy for the trespasses of our ancestors? It does seem like people feel we are to some extent. Perhaps this makes sense, after rigorous and comprehensive reflection: perhaps not. We need be careful to distinguish moral blameworthiness from moral responsibility, more generally, where we feel obliged to make up for the wrongs of our relatives or associates. In the case of original sin – Adam and Eve – they were surely blameworthy via their actions, however, it is unclear we're morally blameworthy in the same way: we did not take that action. Nevertheless, some Christians might suggest that we may be morally tarnished due to the impurity they have passed on to us through the human lineage.

such as works of art or historical antiques. However, it's difficult to conceive of non-conscious objects as morally considerable. Those individuals or philosophies that do assert the intrinsic considerability of non-conscious objects, such as ecosystems or antiques, seem to be, upon investigation, surreptitiously introducing special backgrounds.[31]

In such cases, it might be difficult to expose such special backgrounds or reveal remnant feelings. This is why the need for internal relations is clearer in other cases where special backgrounds aren't as likely to surreptitiously infiltrate our moral judgments. For example, we cannot conceive holding desiccated sticks as morally considerable; in order to do so, we would have to surreptitiously introduce a special background (of consciousness). If a man were to claim desiccated sticks as morally considerable, full-stop, we would consider him either disingenuous, or mad in a way inconceivable to the rest of human beings: that is, inscrutably mad, more so than a person who thought desiccated sticks were in fact highly intelligent creatures who were biding their time before implementing their plans for world domination.

In this way, there are certain constraints as to how we as humans *must* think about morality, in that we cannot conceive of it in any other way.[32]

To couch the discussion so far in Foot's terms: In order for a feature to be a morally relevant feature, certain internal relations must obtain (e.g., the capacity for

[31] Such special backgrounds might include the presumption that ecosystems (bracketing aside all the conscious creatures therein) have a collective consciousness, perhaps in a similar way that neurons while not conscious singularly, together lead to the emergence of a collective consciousness. Some of the insistence by select environmentalists that ecosystems are intrinsically morally considerable may be the connotation of "intrinsic" seems to suggest primary importance whereas "instrumental" seems to suggest a secondary importance.

[32] I am not presenting extensive argument for what these ethical forms of life might be, but I imagine them to be varied: for instance, as previously discussed, that moral considerability requires consciousness; that moral blameworthiness pre-necessitates choice. The claim of ethical forms of life might not be necessarily biologically necessary and could be culturally contingent. For instance, it seems conceivable that paradigm shifts could occur, and morality could be utterly different. Neuroethics and neurolaw, for instance, seem to be eroding the place of free choice in ethics, which could, if we internally accept these findings, lead to a shift in what is an intelligible basis for moral blameworthiness. That ethical forms of life could shift doesn't necessarily undercut my point, however. Given our current state of being, the ethical forms of life do seem set – just as Wittgenstein might suggest that our forms of life, though conceivably variable, could have gone another way. Lastly, in analogy to science, while it is possible that a scientific paradigm might shift, or that a highly corroborated theory that we accept might ultimately turn out to be false, this does not invalidate the paradigm or theory.

choice when it comes to blameworthiness). Foot's example concerns the trivial action of a man clasping and unclasping his hands three times. Without special background, how can this be called a good action? Foot asserts that it only makes sense to talk about this action being good if we imagine some special background: that it fulfills a duty, exhibits a virtue, or, results in some special, presumably positive, effect. Relatedly, a man cannot deem anything harm: for example, taking a bucket of water out of the ocean, or reducing the hairs on one's head to an even number.

Foot's Challenge: Shifting the Burden of Proof

Foot (1959, p. 116) challenges the reader to try to assert certain attitudes about any objects, willy-nilly: "Anyone who feels inclined to say that anything could be counted as an achievement, or as the evil of which people were afraid, or about which they felt dismayed, should just try this out." By this challenge, Foot is pushing the burden of proof upon those who deny that there are internal relations between certain features and moral valuations. To my knowledge, no one has successfully taken her up on this challenge, and provided such an account.

Focusing the discussion upon virtue ethics, Foot (1959, p. 120) asserts that it is "surely clear that moral virtues have must be connected with human good and harm, and that it is quite impossible to call anything you like good or harm. I would argue the same is true, more broadly, of any ethical system or ethical judgment singularly: that is, any ethical judgment must be connected with good or harm. This good or harm might be conceived different ways: good consequences, a flourishing life, preservation of autonomy, avoiding degrading one's own character, realizing one's higher rational nature, producing good consequences, etc. If one presents what she believes to be a counterexample to this assertion, it is likely that the ethical judgment connects back to harm in some sense. For instance, Kantian ethics might say that the immoral action is that prohibited by the categorical imperative, because to do such action would violate one's rationality (as one would be willing a contradiction). What's wrong with doing so? Well we are rational creatures: rationality relates to our higher self and our passions stem from our lower self. We should strive to be rational creatures. This is to our human good.

This is why clasping and unclasping one's hands three times cannot be considered good (though surely it can be *called* good just by stringing the words together). After all, clasping and unclasping does not contribute to any good: it doesn't benefit or harm any subject. Foot presents further examples – taking a bucket of water out of the ocean, or reducing the hairs on a man's head to an even number – to demonstrate that such actions cannot be considered morally bad because they don't constitute harm (though we might dream up special backgrounds to insert into the story in order to introduce harm into the equation).

Denial of Certain Features as Morally Relevant

The second assumption, listed above, challenges moral methodology with the contention that a rational person can simply deny the relevance of certain features as morally relevant. These certain features can be formally represented as factual premises – premises that everyone else takes as counting as evidence towards a moral conclusion. We could imagine such an opponent to be armed with a powerful objection succinctly phrased: "So?" This question is essentially a denial of purported morally relevant features as actually relevant to moral judgment.

For instance, imagine I am trying to convince my interlocutor that consuming factory-farmed meat is immoral. I may provide a battery of facts, which I take to be morally relevant: for instance, eating factory farmed meat causes severe suffering to sentient animals, pollutes the environment, cultivates "super-viruses" immune to antibiotics, is unnatural, destroys family farms, is bad for one's health, and is unnecessary to one's diet. I provide a critical mass of such facts to my formidable opponent. His response to my impassioned diatribe: "So?" What he means by this question is that he takes none of these facts to be evidence in support of the moral conclusion: "One ought not to eat factory-farmed meat."[33]

[33] For our discussion, the "So?" question is narrowly focused in meaning, where it's functioning as a denial of the moral relevance of the factual premises proffered in support of an evaluative conclusion. The "So?" is not to mean that the person simply doesn't care, or care to think, about the morality of the matter. The "So?" question is also not asking for *motivational* reasons why he should care.

In order to reach evaluative conclusions, factual premises are not sufficient to take us there: at least one evaluative premise is necessary. Moreover, it's not clear that, without an evaluative premise, factual premises provide any evidence toward an evaluative conclusion. We are always left with the challenging question, "So?"

Evaluative Premises

No evaluative conclusion can be formally deduced from factual premises; rather, an evaluative premise needs to be present in order for an evaluative conclusion to be entailed. The only way factual premises can be evidentiary, to anyone, is if at least one evaluative premise is present (in argument, it oftentimes it seems these evaluative premises are suppressed). The evidentiary relevance of factual premises is dependent upon the presence of an evaluative premise that validates the facts as morally relevant. The problem is that an interlocutor supposedly needs not accept such an evaluative premise. He may hold that the evaluative premises he *would* accept do not include, as part of the evaluative premise, the facts being proffered as evidentiary. In order for the factual premises to count as evidence, then, the interlocutor needs to first accept the evaluative premise.

In illustration, consider the following argument:

E1. Causing extreme pain to innocent sentient creatures for marginal pleasure for oneself is immoral.
F2. Eating factory-farmed meat causes extreme pain to innocent, sentient creatures.
F3. Eating factory-farmed meat causes only marginal pleasure for oneself.
EC: Eating factory-farmed meat is immoral.

The interlocutor could accept F2 and F3 as factually true, and yet he would not accept F2 or F3 as *morally relevant* toward conclusion EC without first accepting some form of E1. The evaluative premise, E1, refers to descriptive facts and to a moral feature. The objects referred to are creatures that have the properties of being sentient and innocent. "Causing pain to" is a factual description of an action. These factual

descriptions show up in the evaluative premise; they are the features to which the moral evaluation refers.

The question, then, is whether or not the interlocutor can always deny an evaluative premise. If he must accept an evaluative premise, then he will be forced to accept the related factual premises as evidentiary toward the moral conclusion.

I contend that, in certain cases, the interlocutor must accept the evaluative premise. If an individual subscribes to morality, where he believes that moral propositions are objectively and consistently either true or false, then I argue that he must already accept some evaluative premise. If I subscribe to morality, then necessarily I have to accept at least one intelligible basis for morality among the following list (presuming it to be exhaustive): (a) duty (b) consequences (c) autonomy (d) rationality (e) virtue (f) harm/benefit. So if you show that action A is in accordance with (a)-(f), and *not* doing action A is in violation of (a)-(f), then doing action A is moral by the individual's, and any moral subscriber's, own lights. In addition, there are certain morally relevant features which are necessarily relevant to all of these ethical forms of life (a)-(f): for instance, any moral subscription must accept sentience and robust choice as necessary morally relevant features to certain moral evaluations. Subsequently, factual premises that assert these factual features must be accepted as evidentiary toward the evaluative conclusion.

In illustration, consider the well-trodden case of torturing innocent children. It seems that if the interlocutor is a moral subscriber, he must concede the truth of the evaluative premise. Consider the argument represented in the following way:

E1. Causing suffering to innocent, sentient children is immoral because of one or more of the following conditions it fails to satisfy:

a. it violates a duty or
b. it corrupts one's moral character or
c. it makes one a bad human being qua human being or
d. it inhibits one from achieving a flourishing life or
e. it violates the golden rule
f. it does not pass universalization tests
g. it's an action to which one can reasonably object
h. it leads to worse consequences overall (presuming the case is as such)
i. it violates desert

j. it doesn't satisfy (or violates) the function of X's
k. it doesn't express care

2. Sam is a child
3. Sam is sentient.
4. Sam is innocent.

EC: Causing suffering to Sam is immoral.

It seems that if the interlocutor is a moral subscriber, then he must accept premise E1 for at least one of the listed supportive reasons. The interlocutor cannot claim to be a moral subscriber, and at the same time deny all of the supporting clauses listed under E1. If the interlocutor is bound to affirm one or more of the supporting clauses, then the interlocutor must affirm E1. In addition, one or more of the factual premises assert features that are morally relevant to all of those ethical forms of life represented in the supporting clauses. For instance, a consequentialist might not put much ethical stock into desert, so the fact a child is innocent might not be morally relevant to him, in a subscription-relative sense, which means he would deny factual premise 4 as morally relevant. Nevertheless, he would have to admit that premise 3 was relevant and, given E1, serves as evidentiary to conclusion EC.

If the argumentation I am presenting is on-track so far, then we have an example of a moral argument which will *not* break down: the interlocutor must accept some of the factual premises presented as evidentiary, as he affirms the evaluative premise (otherwise he wouldn't be a moral subscriber) and must necessarily accept at least some of the presented features as morally relevant. The thesis, then, that moral arguments are always liable to break down is false.

Let us examine the possibility that an interlocutor denies E1. Imagine that the interlocutor accepts that causing suffering to children is wrong, but denies his moral evaluation is on the basis of the aforementioned features -- innocence, sentience, helplessness, autonomy, or any of the listed or implicit other facts presented as morally relevant. Dumbfounded, we might ask upon features the interlocutor *would* make his moral evaluation that causing suffering to children is wrong, or that causing suffering to *anyone* is wrong. As we saw previously, a person cannot deem any feature, willy-nilly, to be morally relevant. In addition, some features are *necessary* for moral evaluation: for

instance, consciousness is a necessary, though perhaps not sufficient, condition for intrinsic moral considerability. Sentience necessarily implies consciousness. Since consciousness is a necessary MRF for moral considerability, then the interlocutor will have to anchor his moral evaluation, at least in part, upon the evidentiary fact of Sam's consciousness.[34]

Whatever the moral subscription to which the interlocutor subscribes, there will be some overlap of subscription-relative morally relevant features. For instance, consciousness, I have argued, is a necessary MRF for any normative evaluation of intrinsic moral considerability. All moral subscribers must attribute moral considerability to some object, and that object must have, as a necessary feature, consciousness/sentience. Consequently, all moral subscriptions will affirm consciousness of some kind as a morally relevant feature in the evaluative premise. If the factual premise establishes that the object is conscious/sentient, then in virtue of the evaluative premise, which asserts consciousness/sentience as a morally relevant feature, the factual premise now serves as evidence toward the satisfaction of the evaluative conclusion.

Moral Disagreement and Multiple Subscriptions

Though disagreement will persist between different individuals based on their different moral subscription, most moral subscribers accede to more than just one ethical form of life (save for, perhaps, professional moral philosophers). For instance, most subscribers to deontology would still concede that consequences are morally relevant: for example, they would likely concede that employing a moderate yet still coercive torture technique might be morally justified if it saved a million innocent lives from the ravages of a terrorist bomb.[35] Also, most consequentialists will concede that it is only rational to give some deference to the golden rule, and after imagining themselves in the shoes of

[34] Consciousness is a necessary element for sentience, autonomy, rationality, and so forth, which I take to be kinds of instantiations of consciousness.

[35] The oft-cited counterexample of lying to the Nazis in order to save the Anne Frank family from execution is another illustration where consequences present a moral challenge to even the most committed deontologist. Other examples abound, such as killing an innocent in order to save the lives of a significant number of others.

the captured enemy combatant, would at least be motivated to consider subsequently elevating the threshold at which it becomes acceptable to torture for the greater utility.

If the two interlocutors are in reality moral subscribers of multiple ethical colors (not necessarily all to the same vibrancy) then perhaps their ethical disagreement is not as intractable as originally characterized. In reality, it is difficult to imagine consequences as morally irrelevant, just as it is difficult to imagine all deontological considerations as morally irrelevant. Nevertheless, it remains the case that at a given stage of thought, two rational interlocutors can have an intractable, irreconcilable moral disagreement. Nevertheless, it seems patently false that moral arguments may *always* break down. Furthermore, I suggest that the asserted irreconcilability between two interlocutors might be more of an academic point than a worrisome one with which we need concern ourselves (just as we need not concern ourselves with the moral skeptics when considering the moral disagreements people actually have). I believe that moral disagreements, more often than not, derive not from intractable differences of ethical principles, but from disagreement about the facts.[36]

Throughout the several previous sections I have striven to provide philosophical argumentation for the basis of morally relevant features. As mentioned before this argumentation, morally relevant features seem generally agreed upon according to commonsense morality. Given this general agreement regarding what counts as morally relevant and irrelevance, I believe one justificatory approach is similar to that proffered by Rawls: overlapping consensus. This approach would survey the moral subscribers and determine the overlapping core judgments of what is taken to be morally relevant.

Overlapping Consensus

[36] The abortion debate, for instance, is often characterized as an intractable conflict – not as much because there is a difference in moral principles, but because there is a difference in factual assumptions. Both camps support respect for moral persons, but they disagree over whether or not a fetus is in fact a moral person. This moral personhood status hinges upon metaphysical factual assumptions, such as a soul, God's existence and Providence, and so forth.

Though two individuals can disagree regarding which moral paradigm to which they subscribe – say consequentialism rather than deontology – I don't believe they can deny other moral subscriptions as intelligible. Even if the consequentialist only believes that consequences are relevant, he cannot pretend that basing moral determinations on deontology is something he cannot understand, even if he denies it as a proper moral basis. In contrast, the consequentialist cannot find intelligible basing a moral judgment on other considerations: such as clasping one's hands together as a moral wrong, full-stop – that is, without the inclusion of special background. In analogy to grammar, some sentences are false while some sentences are unintelligible. For instance, a Christian might disagree with the literal interpretation of the Nietzschean declaration: "God is dead." Nevertheless, the Christian finds this statement intelligible. The statement, however, "God is alliterative," while grammatical is not literally intelligible. This reiterates the previous discussion of "ethical forms of life": there are certain forms of moral bases we recognize as relevant and others we recognize as irrelevant; this is evident, for one, in our moral concepts and moral grammar.

Not every individual or every culture will agree on what counts as morally relevant, however, even though they might agree on what is morally intelligible. According to Haidt, for instance, there are four "foundations of moral sense" upon which cultures and sub-cultures seem to morally base their values: (1) Aversion to Suffering (2) Reciprocity, Fairness, and Equality (3) Hierarchy, Respect and Duty (4) Purity and Pollution (Sommers, 2005). According to Haidt, in American political culture, liberals recognize only the first two moral bases whereas conservatives recognize all four moral bases.[37] Though liberals do not allegedly recognize the last two bases, they would still recognize them as intelligible.[38]

[37] Other cultures recognize these moral bases: for instance, various cultures in the Middle East emphasize moral judgments on "purity." Objections to these moral judgments from the western world may proceed either from denial of purity as a morally legitimate basis, or from the denial of the presumed facts of "corruption" or impurity, citing them as false. One clear example of this controversy over moral bases is apparent in the debate over female circumcision in some Middle Eastern cultures.

[38] Presumably, this is a contributing factor to the irreconcilability of political debate. Haidt further evaluates liberals as having an "impoverished moral worldview" whereas the worldview of conservatives is more robust. Haidt is an antirealist about the objectivity of morality, but believes that moral claims upon these four groundings should be considered as valid.

To achieve an uncontroversial core of moral unintelligibility, I proffer that an overlapping consensus be established, where this consensus is not one of what moral bases are morally relevant, but rather, what moral bases are morally irrelevant. Presumably, the overlapping consensus of what constitutes morally irrelevant features, after thoughtful examination and rigorous analysis and counterexample, will represent those bases which appear to be morally unintelligible. After establishing this overlapping consensus, any moral judgment that turns out to be based upon a moral basis found in this core of unintelligible bases is deemed noncredible. For instance, after thoughtful examination and rigorous analysis, it is discovered that a moral judgment concerning intrinsic moral considerability is based upon the color of another individual's eyes, this judgment is deemed noncredible, as the basis is morally unintelligible. If one does assert this basis as a proper moral basis, he is making a confusion (which presumably becomes evident after interrogation and counterexample): for instance, the person is unknowingly incorporating a special background. Foot's challenge (1959, p. 89) that anyone who "feels inclined to say that anything could be counted as an achievement, or as the evil of which people were afraid, or about which they felt dismayed, should just try this out," is a challenge based on her recognition that some moral bases are morally unintelligible as moral bases, and moral terms, such as benefits or harms, have necessary internal relations to certain objects, but not *any* objects willy-nilly.

The Happy Torturer

After setting out this account of moral relevancy, and its various means of support, it's elucidative to consider possible counterexamples. One counterexample we might imagine is that of the happy torturer. The happy torturer is not a sadist or masochist, who feels that pain delivers pleasure. The happy torturer just simply judges pain to be intrinsically good, in the same way the rest of us simply judge pleasure to be intrinsically good. Can we not imagine such a moral subscriber? Is this example not intelligible? This seems to take on Foot's challenge: If you think you can connect good to

something willy-nilly – or in this case antithetical – just try it. So here's an example attempting to do just that.

What can we say about the happy torturer? First, we can ask if the happy torturer desires and pursues pain for himself. If he does not, because he himself values pleasure and disvalues pain, then it seems we can gain some traction in that this seems inconsistent: holding pain as intrinsically valuable when experienced by others but not by himself. Furthermore, it would indeed be odd for him to claim that what is intrinsically valuable to others who disvalue the pain or do not seem to be benefited by it.[39] So it must be the case that the happy torturer judges pain to be intrinsically morally good to himself. We might unsavorily imagine that he engages in self-flagellation and other forms of self-torture and experiences solely pain, and no pleasure, and yet he values this pain intrinsically (and, again, it does not lead to pleasure as in the case of the masochist). Is this intelligible to say that he values pain? It appears difficult to imagine. It could be the case that he has inverted pleasure/pain wiring and the feel of a whip lacerating his back is diverted into the pleasure rather than pain receptors, to speak loosely, due to some physiological anomaly. However, this is to divorce the sensation qualia from the valuation of it. In comparison, some might say the sensation of the taste of chocolate is pleasurable, and yet there are people who do not enjoy chocolate. We can imagine, nevertheless, that we are all experiencing relatively the same sensation of tasting chocolate, but some of us find it pleasurable while others find it displeasurable. It seems then that pleasure and pain are more on par with valuing and disvaluing certain sensation qualia. To say, then, that the happy torturer values pain is incorrect: rather, the happy torturer values pleasure, yet he receives pleasure from the sensations from which the rest of us receive pain. If you point out to the happy torturer that while he values qualia T, the kind of varying qualia that is involved in the smorgasbord of torture, the rest of us disvalue qualia T, the happy torturer must realize he is mistaken that torture is good for others. If the happy torturer says, "I realize torture causes pain. I disvalue pain and

[39] I am asserting that the good needs to be valuable or beneficial. It's unclear in this case what the good would be attached to if it is not valuable to the recipient of the alleged good – whether torturer or victim. I do not intend to equate good with valuable, though I do believe in order for something to be intrinsically good it does need to be valuable to someone.

recognize others disvalue pain. Yet torture is intrinsically good." we must shake our heads as truly not having the faintest idea of what he's talking about; his statements seem senseless to us, and presumably we imagine that this set of statements must be senseless to himself as well.

One might alter the example by imagining the torturer not to think pain is intrinsically good, but holds that pain is instrumentally good in that it always leads to benefit. Furthermore, he may torture people thinking that pain is not intrinsically bad at all; in addition, it is instrumentally good. If it claimed that the "instrumental torturer" doesn't believe pain is bad, believing rather that pain is neutral – neither intrinsically good nor bad – then I would assert this claim would be as intelligible as believing pain to be good, and for the same reasons. If the instrumental torturer believes pain to be intrinsically bad while concurrently being an *instrumental good*, then this is in fact intelligible to us, though it may turn out to be factually false: that the torture inflicted does not outstrip the benefits.

Conclusion

In this first chapter, I have delineated the basic framework of this dissertation project. The goal of this project essentially is to fortify the moral methodology of wide reflective equilibrium, with special attention to developing the Rawls' filtration process, which determines credible from noncredible intuitions. In order to determine the credibility of moral intuitions in a substantive and nontrivial way, some initial account needs to be provided and defended in regard to what features are morally relevant or irrelevant. In the latter portion of this chapter, I have attempted to provide such an account based on four approaches: (1) ordinary morality (2) ethical forms of life (3) shifting the burden of proof (4) overlapping consensus. All four approaches are interrelated and interdependent. Having established an account of moral relevance that is hopefully satisfactory for our purposes, we can move, in the following chapters, to consider moral judgments and their place in wide reflective equilibrium. In the next chapter, I will sketch out the method of reflective equilibrium, and exposit further the

basic elements of the methodology, including its starting assumptions. In expositing this methodology, I will argue that this methodology is a valid and promising one.

Chapter 2: "Methodologies of Reflective Equilibria"

Introduction

In this chapter, I will present and defend the method of wide reflective equilibrium (MWRE) as a moral methodology. I begin by delineating the method of narrow reflective equilibrium (MNRE) and I then consider objections that have been leveled against it. I move on to show how expanding MNRE to MWRE by the inclusion of background theories enables MWRE to overcome the inadequacies of its predecessor. As part of my defense of the method of reflective equilibrium (MRE), I will underscore the importance of bringing within the scope of MRE, not only scientific background theories, but also the considered moral judgments and moral principles of others.

Preliminary Points

Before I delve into the exposition of the method of reflective equilibrium, a few preliminary points should be established. In the following subsections, I will provide a brief delineation of these items, leaving further development, as necessary, until later in this and the following chapter.

A. Initial Judgments versus Considered Judgments

Initial judgments refer to those moral intuitions that have not been quality-checked, via the filtration procedure, to determine if they are based upon any error-disposed conditions. For instance, suppose an individual has the initial judgment that "In self-defense, I can strike this man who is stabbing me with a metal rod." That individual should consider if this moral intuition is based upon any error-disposed conditions, such as a lack of conceptual clarity. In this case, the individual having the initial judgment lacks conceptual clarity if the man is a doctor who is trying to assist him, the patient, and the sharp metal rod is a needle filled with a life-saving vaccine.

Considered judgments, then, are those initial moral judgments that have been quality-checked, in that they have been determined to be free of error-disposed

conditions – such as mistaken beliefs or lack of conceptual clarity – via the filtration process.

B. Initial Credibility and Moral Objectivity

Three interrelated points need to be made at the outset regarding the status of initial judgments in the context of the method of reflective equilibrium. First, a necessary starting assumption of MRE is moral objectivity of some sort. Second, a further assumption is that initial judgments are initially credible in that they provide us some kind of access to this objective moral truth. Third, in presentation of these two former points, some defense should be provided to show that these staring assumptions are plausible and reasonable.

1. Moral Objectivity

The starting assumption of moral objectivity needn't be married to any precise conception of moral objectivity. A general conception of moral objectivity is sufficient, where moral propositions are consistently either correct or incorrect. This conception denies moral skepticism, but needn't presume moral realism or that there is some kind of "mind-independence" to moral propositions. For instance, on one antirealist interpretation of a Humean sentiment-based ethics, a moral proposition is objectively correct if it comports with the human sentiments, tempered by practical reason, and is objectively incorrect to the degree it departs from the human sentiments. Likewise, if one is a Kantian, one finds rational reasons as to what ought to be done and what ought not to be done. These propositions are objectively normative in that they are action-guiding as they provide reasons for us to do certain actions and refrain from doing other actions. Moral objectivity can even come with certain non-cognitivist views, which allow moral judgments to be either correct or incorrect without being true or false, where correctness and incorrectness share certain formal properties with truth or falsity.[40] For instance, a judge's verdict could be considered correct or incorrect, without being true or false,

[40] For example, that no moral judgment is both correct and incorrect, that either any moral judgment is either correct or incorrect, or that they can be neither correct nor incorrect for reasons similar to why judgments of the sort that are true or false can admit of truth-value gaps, and so forth. These formal properties were helpfully suggested by Norman Dahl.

depending on whether it comports with or departs from law and legal precedents; moreover, we would consider a judge to have normative reasons, where he ought to deliver correct rather than incorrect verdicts.[41] Normativity, in this general way, provides genuine practical force in that there are objective reasons to act in certain ways while not in others. When I refer to "truth" in this dissertation project, I am referring to the degree moral judgments approach correctness in this general normative sense.

2. Initial Credibility

We take our sense experiences seriously as providing us some sort of access to an objective world, and we further assume these sense experiences to be consistently either correct or incorrect.[42] In a similar way, initial judgments are initially credible if they provide some access to moral objectivity. Moral intuitions needn't be on par with direct sense observations, but they need to reliably provide some sort of access to moral objectivity, where initial judgments would be credible if the moral evaluations included in their content related to what is out there in the world. For example, my initial judgment "killing my neighbor's child is wrong" has initial credibility if the reason I have this intuition is due to features of the world where it would be, at least, incorrect to act in such a way as to kill my neighbor's child. If we assume there is some kind of objective moral truth, and we can gain access to it, it seems plausible to think that this access will come at least in part through our intuitions.

All moral theories that acknowledge moral propositions to be objectively true or false, or correct or incorrect, end up taking certain moral intuitions to have at least initial credibility.[43] Neo-Kantian ethicists are troubled by the Inquiring Murderer; utilitarianism by the Framed Innocent Man; virtue ethics by the Mafioso; and so forth. These moral

[41] It's important to distinguish objective statements from (individually or culturally) subjective statements. A statement expresses a subjective judgment if one person could affirm the same statement which the other denies, and yet they both could be correct. A statement expresses an objective judgment if for any two people if one of them were to affirm it, and the other deny it, at least one of them has to be mistaken (presuming it is a matter where denials or affirmations of it can be true or false).

[42] For example, a man who sees flying pink elephants seems to be having a sense-experience that is incorrect.

[43] By "moral subscription" I mean any moral theory or moral outlook concerning desert, duty, rights, consequences, preferences, liberties, virtues, and so forth.

theories recognize that even if we have a map of moral principles, without the compass of moral intuitions, our ability to navigate the moral landscape would be lost. Principles without intuitions are empty, and intuitions without principles are blind.[44]

The starting assumption of initial credibility of our intuitions seems plausible. Over the human history of moral thought, it seems reasonable that we have made progress in our contemplation of moral matters. Furthermore, initial credibility fits our normal beliefs and practices. In our ordinary lives, liberated from extreme skepticism, if we were not to recognize our moral intuitions as carrying some initial credibility, then the moral dimension of our world would disappear and become inscrutable.

The question we are left with, then, is this: Granted that there is objectivity of some kind, what methodology seems the best candidate to discover moral features of our world? I will argue that, granted moral objectivity, MWRE is the best moral methodology. In particular, I will argue that we should grant MWRE its starting assumption of objectivity, given its relevant similarities to scientific methodology.

C. The Analogy to Scientific Methodology

One initial argument for taking MWRE seriously, including its starting assumptions of moral objectivity and initial credibility, derives from its close relation to scientific methodology.

Throughout this chapter and the next, I will explore the analogy between the method of reflective equilibrium, especially MWRE, and scientific methodology. MWRE parallels scientific methodology in both ontology and structure: moral intuitions parallel experimental/observational data; moral principles parallel scientific hypotheses; MRE background theories parallel scientific background theories. MWRE also parallels scientific methodology in process: intuition generation conditions parallel the experimental/observational conditions necessary for the generation of credible data; credible data is systematized into hypotheses, corroborates or discredits hypotheses, and

[44] Any poignancy to this statement is attributable to Kant (1998, B 76), who states in the Critique of Pure Reason, "Thoughts without content are empty, intuitions without concepts are blind." Another quote that is also often attributed to him states: "Experience without theory is blind, but theory without experience is mere intellectual play."

is itself judged in coherence with corroborated hypotheses; lastly, background theories, in both methodologies, bear upon both intuitions/data and hypotheses/principles, by their conflict or coherence.

The scientific method is accepted as a valid methodology to arrive at correct empirical data and correct corresponding hypotheses and theories. For similar reasons, I will argue that MWRE should be accepted as a valid methodology to arrive at correct moral judgments and correct corresponding moral principles and theories.

D. The Assumption of Rationality

Scientific methodology assumes a conception of rationality; MRE will be assuming a similar conception. Though science has no settled definition of rationality, the scientific methodology generally characterizes rationality, in part, as a criterion that restricts a scientist from being biased toward a pet hypothesis in the face of a critical mass of opposing counterevidence: a scientist should be willing to submit her hypothesis to critical tests, and to abandon it if it confronts a critical mass of counterevidence.[45]

This invites the problem of Duhem's (1954) thesis concerning falsifiability. If a scientist is wedded to her hypothesis, she may end up discarding all data sets that fail to cohere with it. She needn't even discard the anomalous data set, but can always discard some other hypothesis that is part of the bundle of hypotheses/theories involved in the test. Rationality prohibits a scientist from abandoning hypotheses in the bundle that have been highly corroborated via independent tests. For instance, if the scientist would rather abandon the highly corroborated hypothesis known as the law of the conservation of energy, rather than abandon her pet hypothesis, this would be irrational, given that the law of the conservation of energy is far more corroborated, through independent tests, than her pet hypothesis.

Imre Lakatos (1970) points out that it is not necessarily irrational for scientists involved in a research program to maintain their theoretical core of theories from falsification attempts by shielding it via auxiliary hypotheses that can take the fall in their

[45] What counts as a "critical mass" is itself a general concept where no precise line can be drawn. Scientific methodology would agree, however, on various outliers: cases in which a scientist is either too ready or too recalcitrant to abandon a hypothesis in the face of counterevidence.

stead. Lakatos acknowledges, however, that research programs can be progressive or degenerative: if a research program is degenerative it lacks growth, increases its protective belt of auxiliary theories, and/or does not lead to novel facts (pp. 176-177).

E. Reasonability

A component of the rationality assumption, that might deserve brief mention, I am calling reasonability. Reasonability requires acceptance of non-skeptical metaphysical positions, such as the denial of solipsism. Reasonability also requires acknowledgement of the validity of the scientific method, and, contrarily, the rejection of invalid methods, and non-substantiated claims of pseudoscience.[46]

F. Three Forms of Credibility

For the purposes of clarity, it's important to distinguish between three types of credibility. The first type is initial credibility, which is granted to initial judgments as a starting assumption of MRE: this credibility is granted to moral intuitions that haven't been, as of yet, subjected to the filtration process. The second type of credibility is preliminary credibility, which is attributed to considered judgments: those initial judgments that have survived the filtration process.[47] Determining preliminary credibility is a main focus of this dissertation project; and when I employ the term "credibility" it can be assumed I am referring to preliminary credibility.

---

[46] This "reasonability" criterion will exclude individuals who deny that certain scientific theories are substantiated, such as evolution, and accept pseudoscientific theories. A creationist might deny certain theory filters, such as genetic relatedness between human beings and primates, which could relate to intuitions about sentience of primates. While I acknowledge there may be no fine distinction between science and pseudoscience, I believe the broader distinction is accepted and understood. Reasonability extends to background theories as well, which may be mischaracterized by unreasonable individuals who are biased in terms of politics, religion, or ideology where they deny highly corroborated, though not decisive, theories, while subscribing to uncorroborated theories that cohere with their beliefs.

[47] The sense of "preliminary" I am using I take to be similar to the way the term is used in other fields: for instance, in describing "preliminary" studies concerning the side-effects of certain pharmaceuticals, the preliminary studies establish a certain data set, which is credible in that the data were generated under proper experimental conditions; nevertheless, the data set has to be corroborated by more extensive research. For example, it could turn out that the preliminary data set, though vetted as credible, proves to be an outlier to the data sets generated upon more extensive research. In some cases, this preliminary data set would be discarded -- even though the set was originally deemed credible – upon the arrival of more extensive research that delivers further data sets that are larger, and incompatible with the findings of the preliminary set.

Norman Daniels (1996c) explains that some of the credibility that considered judgments gain in MRE derives from *systemization* in reflective equilibrium. I refer to this third type of credibility as "systematized credibility." Systematized credibility can either be narrow or wide, depending on whether the CJs cohere with just moral principles, as in MNRE, or with BTs in addition, as in the case of MWRE.[48] In the case of MWRE, additional credibility is attributable to *wide* systematized considered judgments, given the systematized coherence between the tripled rather than doubled set.

Moral judgments gain credibility through the filtration process and narrow or wide systematization. Credibility, then, is a characteristic moral judgments gain not through their content but through process: considered judgments have survived the trial by fire of the filtration procedure; systematized considered judgments have survived a mutual adjustment via reflective equilibrium. Despite the credibility that systematized CJs gain via coherence with moral principles and, in the case of MWRE, background theories, moral judgments never gain any special epistemic privilege after passing the filtration procedure and surviving systemization. CJs and SCJs, both, are always provisional, and subject to exclusion if an RE set undergoes revision.

Having established these preliminary points, I will now move to the body of the chapter, and discuss considered judgments and their place in the method of reflective equilibrium. From there, I will exhibit the method of narrow reflective equilibrium, and examine both its strengths and limitations.

Considered Judgments

Human beings have moral intuitions in regard to actions, dispositions, character traits, and the like. The method of reflective equilibrium begins with such initial moral

[48] A fourth type of credibility, considered later, is broad systematized credibility (BSC), which is an increased credibility attained over multiple RE sets, when an individual's SCJs cohere with not only her own moral principles (and possibly background theories), but relatively identical SCJs are found to be coherent to the moral principles and background theories of other individuals. This is similar to a scientist discovering the empirical data she's generated is relatively identical to the empirical data generated by other scientists, that it is coherent with foreign hypotheses, or finding that the foreign data coheres with her own data and her own hypotheses (and possibly her data cohering with the other hypotheses of foreign scientists).

intuitions. They serve as provisional starting points of moral theory construction. However, not all of an individual's intuitions are to be granted provisional status: some intuitions may lack credibility. An intuition lacks credibility when the intuition is likely to have been generated under error-disposed conditions. Rawls (1971, p. 176) mentions a few of these conditions: immoderate emotionality, self-interested bias, and ignorance of relevant facts. The first task of moral theory construction, via MRE, is filtering error-disposed intuitions from credible intuitions. This filtration process is a crucial mechanism in MRE, which I will discuss in much more detail in the next chapter.

Considered judgments are asserted by MRE to be our provisional starting points. However, these considered judgments don't enjoy any special epistemic status other than that they are deemed preliminarily credible, because they have been vetted by the filtration process as not disposed to error. In this way, considered judgments do not claim the same eminence as many traditional versions of ethical intuitionism ascribe to intuitions: intuitions are not, for example, deemed "self-evident" or considered direct apprehensions of moral features of the objective world.[49] The term "credibility" refers to epistemic rather than substantive moral concerns. The claim is not that the source of each intuition somehow imbues each respective intuition with moral normativity.[50] The claim is that intuitions are provisional data points of normativity, which means they are taken to be initially credible, and that those provisional data points need to be quality-checked in relation to their generation conditions, after which an intuition is deemed to have preliminary credibility. It is epistemic credibility that is at issue concerning this

[49] W. D. Ross characterizes intuitions as self-evident (1930, p. 29); G. E. Moore treats them as direct apprehensions of moral features of the world (1903, p. 148).

[50] Normativity is a substantive moral property. Credibility, in contrast, is an epistemic property. In illustration, consider three brief examples. Divine Command Theorists might claim that if a prescription "Thou Shalt not X" is commanded by God, the prescription, whatever X turns out to be, becomes imbued with normativity. In contrast, some sociobiologists have claimed that the biological origin of a moral prescription would automatically undermine its moral normativity. In a third example, we might imagine that a hypnotist might dispose a subject to feel that doing X was immoral. Depending on what X is (e.g., that clapping one's hands is immoral; or that striking a baby is immoral), whether or not the prescription is actually morally normative is contingent; given the etiology of the intuition, though, we can dismiss the resultant intuition, whatever it turns out to be, as noncredible, even though it may happen to be actually normative. In this project, I take no stance on normativity itself, other than to presume as a starting assumption that initial judgments are initially credible assertions of moral normativity: that is, that they should at least initially be taken as "true" when they prescribe "One ought X" or "One ought not X."

filtration process, not normativity – though if an intuition loses credibility, it likewise loses any claim to be something that should be taken to be normative. To determine credibility, initial intuitions must survive the filtration process, which filters out error-disposed judgments from the set of initial judgments, resulting in set which Rawls calls considered judgments.

The filtration in MRE functions similarly to data-gathering in scientific practice: if data is generated under experimental or observational conditions that are error-disposed, scientists will typically discard those sets of error-disposed data from the body of data involved in hypotheses corroboration and theory construction. For example, if a scientist is proposing to test the relationship between pressure and volume of a monatomic gas when the temperature is maintained at a constant; the scientist must ensure that the temperature was in fact maintained during the experiment and did not fluctuate. If the scientist discovers, after the experimental data is generated, that the temperature may have fluctuated (for instance, he might find out the electric thermometer was malfunctioning), he should filter out that data set from the accepted sets for this experiment. This filtering of initial data in science, given experimental conditions, is similar to filtration in MRE: Initial intuitions are to be considered data; yet if we find that an initial judgment was generated under error-disposed conditions, that initial judgment is to be discarded from the set of credible data. Only considered judgments, vetted by the filtration process, are to be accepted as provisional starting points in MRE.

I will further discuss the status of considered judgments later in the chapter. There, I will consider Daniels' (1996a) claim that considered judgments in MWRE seem to require less initial status than they do in MNRE, given the inclusion of background theories in the latter methodology. One last point regarding considered judgments and their relationship to credible data in science: both might be considered to be evidential, however indirectly, of what reality is like. For instance, Daniels (1996a) states that while considered judgments are neither self-evident, nor direct apprehensions of moral features of the world, they might be taken as "evidence" of moral objectivity of some kind) I will develop this topic in the sub-section entitled "credibility and moral objectivity."

The Method of Narrow Reflective Equilibrium: An Overview

John Rawls (1971) proffers a normative methodology based on the coherence between judgments and principles, now known as "narrow reflective equilibrium." As characterized by Daniels, narrow reflective equilibrium is a system comprised of a doubled set: considered judgments and moral principles.[51] I will refer to this doubled set of considered judgments and moral principles as an "NRE set." This is to be distinguished from the MNRE which denotes the method of narrow reflective equilibrium as a moral methodology. Similarly, as I will expound later, a "WRE set" refers to the set of considered judgments and moral principles as well as the addition of background theories. Individuals may differ in what NRE or WRE sets they subscribe to, and some reflective equilibrium sets will be in greater equilibrium than others, enjoying a stronger coherence than other sets.[52]

MNRE begins with a set of initial judgments, which are then pruned of error-disposed judgments via the filtration process. The filtration process, the focus of the next chapter, appeals to theories -- nonmoral and normative, scientific and metaphysical -- to establish "relevant cognitive conditions."[53] Relevant cognitive conditions are those conditions in which an initial judgment arises that allow that intuition to be deemed credible. In contrast, error-disposed conditions are those conditions that determine an initial judgment to be noncredible. For instance, psychology might inform us that when

[51] An NRE set denotes the set of CJs and MPs that are in reflective equilibrium with one another, whereas MNRE denotes the methodology. An individual at any given moment might have a certain set of CJs which are, or are not, in equilibrium with certain MPs. In this way, we can refer to two or more individuals who may have different narrow reflective equilibrium sets (NRE), but both are applying the same moral methodology of narrow reflective equilibrium (MNRE). This is similar to the distinction between scientific methodology, and a particular set of scientific data in equilibrium with a set of scientific hypotheses. There might be competing sets of scientific data and hypotheses relating to some empirical phenomenon.

[52] This is similar to the practice of scientists, who sometimes differ in what data sets and hypotheses they accept. Some data set that is accepted by one scientist might be bracketed as anomalous by another. In this way, two scientific hypotheses may be in competition, where scientists strive to show how their hypothesis coheres with the data better than another.

[53] Tom Regan defines "relevant cognitive conditions" as those conditions necessary for generating intuitions to be credible. I will provide a detailed explication of these conditions in the next chapter, in specific contrast to error-disposed conditions. For now, it should suffice to say that RCCs are those conditions that are not error-disposed. For instance, conditions where the subject has the correct facts, clear concepts, and basic rational capacities (1983, ch. 4.2).

individuals are severely angry in certain situational contexts, they are more likely to generate initial judgments that run afoul of the facts. A driver who is rear-ended by another car, for instance, might generate an initial moral judgment that presumes the offense was intentional when, upon empirical examination, it turns out such collisions are almost always accidental. In such cases, the initial judgments that arise under identified error-disposed conditions should be filtered from the set of considered judgments.[54]

The initial judgments that survive filtration are called considered judgments. Abstractions are then made from these considered judgments to form moral principles: the principles represent a systematization of certain subsets of considered judgments.[55] If principles have already been systematized from subsets of considered judgments, subsequent considered judgments serve as tests of the principles, which either corroborate or discredit the principles. If there is conflict between some set of considered judgments and some set of principles, either the conflicting considered judgments or conflicting principles must be discarded or revised, or both.[56]

---

[54] Error-disposed conditions will be explicated and justified in the next chapter. EDCs in MWRE are analogous to those in scientific methodology, where certain conditions must be met in order to ensure fidelity of generated data. For instance, in psychology, double-blind studies are an ideal goal of experimental setup, in order to avoid biasing conditions that might undermine the credibility of the results. In chemistry, temperature, pressure, and volume are parameters that need to be controlled in order to ensure credible data. If an experiment is performed, which generates a set of data, and it is later discovered significant error-disposed conditions were not excluded in the experimental setup, that generated data is filtered out of the data set for that experiment type.

[55] Principles can also be prior to considered judgments. Social cultivation from youth to adulthood may instill moral principles in individuals. Considered judgments may then originate after the principles, and serve to check the principles. A concern might arise, here, that any subsequent considered judgments will be "theory-laden." This will prove more of a problem for MNRE than MWRE. I will tend to this concern later in the chapter.

[56] I will provide an example, in illustration, later in this section. For now, it might suffice to mention that that Philippa Foot provides a good example of what appears to be MNRE in her article, "The Problem of Abortion and the Doctrine of Double Effect" (1967, pp. 5-15). In this article, she argues that our moral intuitions cohere better with her doctrine of positive and negative duties than the doctrine of double-effect, itself. In order to show this, she considers a diverse sampling of hypothetical examples, examining our common shared intuitions about the right and wrong action in each thought-experiment. Foot finds that her principle – a doctrine of positive and negative duties – better coheres with our intuitions in such cases than the doctrine of double-effect. Foot also considers other principles, such as "The Catholic doctrine on abortion," which prohibits killing a fetus even if the fetus is already going to die, and killing the fetus is the only way to save the mother. Foot says this doctrine "must here conflict with that of most reasonable men" (1977, p. 30). That is, this principle conflicts with the moral intuitions, or corroborated principles, of most reasonable people, and therefore should be revised or discarded.

As previously mentioned, scientific methodology and MRE assume a similar conception of rationality. This conception of rationality demands that an individual be ready to submit her hypotheses to severe and critical tests, and furthermore be willing to abandon or revise a hypothesis if it fails to pass such tests. In some cases, the appropriate revision will be clear: for example, regarding MRE, if a set of principles perfectly systematize all considered judgments save one, the presumption will be that the singular considered judgment should be excluded from the credible data set. This is similar in cases of science when a law accommodates all data points, except for one; usually that one anomalous data point will be discarded as an outlier (and presumed to be inaccurate). However, there is still the possibility that this singular anomalous CJ is so compelling and fundamental, whereas the other mass of CJs and principles seem less compelling.[57] In this case, the strength of this one CJ might cause us to question the conflicting principle.

This latter possibility should not prove indicative of any critical problem with MRE as a moral methodology, since a similar situation obtains in science: if a hypothesis cannot accommodate a data point that is highly corroborated, then that hypothesis is subject to revision or rejection. Examples in the history of science also illustrate cases in which an anomalous data set, which was initially discarded, was not inaccurate, but rather a legitimate data point, whereas the hypothesis or law it conflicted with was false.[58] This exemplifies a practical difficulty with MRE as well as scientific methodology. Such examples show how anomalous data may be sometimes bracketed as inaccurate, only later to be determined accurate. I use the term "bracket" to indicate that this anomalous data is only provisionally set aside, and is not actually rejected once and for all. Anomalous but ostensibly credible data always need to be considered and

[57] For example, the considered judgment "torturing children for fun is wrong," while not being a "self-evident truth" certainly is a very compelling intuition. We can imagine that if the rest of our CJs, and the principles these CJs supported, were less compelling, we might abandon both the principles and respective CJs, and start anew with principles that cohered with this CJ. This doesn't mean that such an intuition is foundational; all CJs are provisional starting points; as such, every CJ is subject to revision. The "torturing children for fun is wrong" considered judgment might seem so steadfast because it enjoys a high degree of systematized credibility, both widely and broadly. I will explain these expanded notions of credibility later in the chapter.

[58] The famous Eddington experiment in 1919, which tested two predictions relating to the bending of light near the edge of the sun, showed that the predictions calculated from Albert Einstein's General Relativity against those of Newtonian mechanics, provided results corroborating the former rather than the latter. These results, despite the reports of many textbooks, were not immediately accepted by all scientists.

reconsidered if error-disposed conditions are not evident or likely; only the identification of error-disposed conditions justify outright rejection of generated data.

There are a few ways to reconcile anomalous data: one way is to presume the anomalous data is erroneous, even though the error-disposed condition has not as of yet been identified.[59] Another way to reconcile consistent anomalous data is to accept the hypothesis as not yet perfected, or to view the hypothesis as perhaps only a model that suffices as the best for now until a better hypothesis comes along.

In illustration of how MNRE functions, consider the following example. Imagine that Edward has an initial moral judgment that eating factory-farmed meat is morally permissible. However, Edward also holds the moral principle: "One ought to do no harm unless, at the very least, the gain is of considerable moral significance."[60] I will simply refer to this as the harm principle. Edward realizes that eating factory-farmed meat causes harm to animals: it causes them suffering and death. Edward also realizes that his enjoyment of eating factory-farmed meat is not of considerable moral significance. In this case, his initial judgment conflicts with his harm principle.

What Edward might first do upon becoming aware of this possible tension between his initial moral judgment and principle is to submit his initial judgment to the filtration process: Is there reason to think his initial judgment arises under error-disposed conditions? It might turn out that his initial judgment does arise under error-disposed conditions: For instance, perhaps one error-disposed condition might be his factually mistaken presumption that animals are not sentient to any significant degree. That is, various theories included in the filtration procedure, such as cognitive science and

[59] For example, perhaps the anomalous data consistently results from a methodological mistake, or from perhaps a material flaw in the experimental setup. A methodological mistake might be that the study was not double-blind, and thereby failed to exclude bias or undue influence in data generation. In illustration of a material flaw in an experimental setup, imagine the light from a microscope's illumination introduces energy into what is presumed a closed system, which affects the average movement of some photosynthesis-metabolizing protozoan organism, whose average movement is being measured.

[60] The term "considerable morally significance" might be contentious, here, as it already seems to be importing some moral presumptions. Moral determination will never occur in a vacuum without the presumptions of some moral principles or normative laws, just as scientific methodology does not occur in a vacuum, but must presume scientific laws and theories. For the purposes of the example, we can assume that "considerable moral significance" is broadly defined, and acceptable to all moral subscriptions. I will presume that "doing harm" is morally bad on any moral subscription – that is, any moral theory or principle one subscribes to – absent any countervailing conditions.

neurophysiology, can reveal that Edward's initial judgment was based upon an incorrect factual belief: namely concerning the sentience of animals. If this is the primary basis for Edward's intuiting that it is morally permissible to eat factory-farmed meat, Edward's initial intuition should be discarded.

Consider a slightly alternative story, where we imagine that, after diligent search for error-disposed conditions concerning Edward's initial judgment, no error-disposed conditions are found. Resultantly, Edwards's initial judgment is now deemed a considered judgment. We then must weigh the considered judgment against the strength of the principle. If the harm principle were highly corroborated by other considered judgments, then rationality would prescribe Edward abandon his singular anomalous considered judgment.[61] Edward would have additional reason to do this if he found his anomalous considered judgment conflicted with not only his harm principle but with other considered judgments of his: for instance, if he morally intuited, under error-free conditions, that domesticated animals ought not to be eaten, exploited, or harmed, and that they were no different than farm animals.

The Narrowness of Narrow Reflective Equilibrium: Criticisms of MNRE

Various criticisms have been leveled at MNRE. Richard Brandt (1979, p. 22) characterizes narrow reflective equilibrium as a mere "systemization of prejudices." The criticism rests on the fact that considered judgments are abstracted to form principles, and principles check considered judgments. Not only does this type of justification procedure seem circular, more critically, MNRE seems to lack any independent justificatory grounding. For these reasons, MNRE is often caricatured as merely a more sophisticated, non-rigorous version of intuitionism.

Daniels concurs with Brandt, though not nearly to the same degree, that MNRE is problematic in that it lacks "traction." His allegation is that MNRE is simply a mutual

[61] The corroborative strength of the harm principle might rest upon the basis not just of Edward's own considered judgments, but the considered judgments of other individuals as well. In addition, the corroboration of the principle would be even stronger, similar to the case of other's CJs, if Edward's harm principle cohered with not only his other moral principles but the moral principles of others as well.

matching and adjustment between considered judgments and principles (which are abstracted from the considered judgments). This means the CJs and the principles may be in reflective equilibrium with each other, and yet lack any further grounding that independently substantiates either set. That is, a set of principles may perfectly reflect a set of CJs, and vice versa; however, this concurrence could still be trivial and arbitrary.

These criticisms underestimate the justificatory power of MNRE, however. First, it would seem to lend some *prima facie* justification if *new* and original considered judgments matched our set of established principles; this is a way in which MNRE could have said to have "predictive power," in that it was able to coherently accommodate future considered judgments.[62] This help could serve as an initial, if not decisive, response against the charge that MNRE is merely a *post hoc* abstraction of considered judgments that lacks any traction.[63] It would seem to recommend *some* credibility to a moral principle in an NRE set if it turned out to be reliably predictive of future moral judgments generated by an individual – or, even more so, generated by others who didn't share the moral principle. If a moral principle turned out predictive in this way, in that it cohered with such future considered judgments not yet generated, this would avoid the change that the individual who held the moral principle was merely generating subsequent moral judgments through a theory-laden prism of his moral outlook.

A second initial resource for MNRE relates to those criteria that characterize good scientific methodology. For instance, MNRE can find some substantiation by appeal to the principle of parsimony: namely, an NRE set that contains fewer moral principles that sufficiently explain the data is, all else being equal, superior to a competing set of principles that contains more principles needed to explain the same data.[64] This is a

---

[62] This is similar to scientific practice, where a hypothesis is abstracted from data, then new data is generated to test whether that constructed hypothesis can accommodate this new data. If the hypothesis can, in fact, accommodate the new data set, then it gains some evidential support from this corroboration. We might imagine the Foot's doctrine of negative and positive duties would be more highly corroborated if it cohered with new thought-experiments she had not considered, especially in novel contexts.

[63] This is not to suggest that *post hoc* objections are innately objectionable; many scientific explanations are *post hoc* and yet still good explanations. For instance, the theory of continental drift, while a *post hoc* explanation, is a theory with better explanatory power than competing theories.

[64] The justification of the principle of parsimony is that all assumptions introduce possibilities for error: therefore, the fewer assumptions, the fewer possibilities for error. A contrary principle that might be considered asserts that the fewer the assumptions, the broader their scope, and, consequently, the greater

fundamental principle of scientific methodology: all else being equal, the explanatory hypothesis that is simpler is the superior explanation, given its wider explanatory power.

A classic scientific example of the principle of parsimony, often referred to as "Ockham's razor," is the competition between Ptolemy's astronomical model versus a heliocentric astronomical model, as first proposed by Copernicus, and improved upon later by Galileo. The Ptolemaic model explained the retrograde motion of Mercury, relative to Venus, by introducing epicycles within Mercury's orbit. Far simpler, the heliocentric model explained this same retrograde motion by replacing the Earth with the sun, as the system's center, while opting for elliptical rather than circular orbits. In addition, the heliocentric model eliminated the supposition of crystalline spheres in which the planets were said to be embedded. We can understand this in the context of MRE by considering the following case: Imagine that there were several competing NRE sets, where the principle to CJ ratio in each of these competing sets was 2:1: that is, one principle cohered with every two initial judgments. These NRE sets would be very crowded. Consider that another NRE set was competing along with these other inter-competing sets; however this set was comprised of one singular principle coherent with every single considered judgment, individually and cross-culturally.[65] Presumably, we would consider the latter NRE set much more substantiated than the former crowded NRE set. There might, of course, still remain two or more NRE sets that equally satisfy the parsimony principle; this challenge can be ameliorated later by the introduction of other similar scientific kinds of criteria, such as fecundity (as well as by the inclusion of background theories, when we later expand MNRE to MWRE). By appealing to prediction and parsimony, MNRE reveals that it has resources similar to those available to scientific methodology. I will explore this parallel further later in the chapter.

In addition to these prediction and parsimony, a robust filtration process, fortified by theory filters, provides substantial resources to MNRE to defend itself not only against

---

chance for error. This seems to illustrate the trade-off between qualitative coherence versus quantitative coherence. One can usually achieve greater qualitative coherence by increasing quantitative entities, such as principles; however, this greater qualitative coherence diminishes the justificatory power of the principles, and likewise their validation. This question, while interesting, I shall not pursue further here, but leave it to the investigations of the philosophy of science.

[65] Presuming that it turned out, after filtration, that no considered judgments conflicted with each other.

Brandt's charge that MNRE is nothing more than a systematization of prejudices, and also against the more general objection that MNRE lacks traction. MNRE is "grounded" if we include independently substantiated theories in the filtration process: this inclusion can provide significant justificatory support to MNRE in that the inclusion of these theories in the filtration process can provide an independent basis upon which to prune error-disposed initial judgments – intuitions that run afoul of the facts, for instance – from the set of considered judgments.[66]

In consideration of Brandt's objection, imagine the case of a white supremacist who generates initial moral judgments that evaluate non-whites as having lesser moral considerability than whites. Further presume that these initial judgments pass filtration and can be thereby deemed considered judgments. These CJs turn out to match the white supremacist's moral principles, and vice versa. If such a case were possible, it would appear MNRE might just be, as Brandt characterizes, "a systemization of prejudices."

Such a scenario seems facile, however. Though easy to describe in words, conceiving such a case that passes filtration is relatively difficult. The racist's initial judgments, to be credible, must be based upon some morally relevant feature, and must not have generated under error-disposed conditions. For instance, we would need to imagine that the supremacist's initial judgments were based upon accurate understanding of relevant facts and concepts. If the racist's initial moral judgment – that nonwhites are less morally considerable than whites – is based upon the factual mistake that nonwhites are cognitively inferior to whites, then (given empirical substantiation included as theory filters in FP) this initial moral judgment would not survive filtration.

That is, if the white supremacist's prejudice against non-whites is because he thinks all non-whites are innately less intelligent, we could utilize empirical tests from psychology, sociology, and cognitive sciences, in order to establish the racist's initial moral judgments to be noncredible given their basis on false propositions. The white supremacist is predicating his initial judgments on falsehoods. Empirical sciences,

[66] For example, if a racist is prejudiced against Blacks because he thinks them innately less intelligent, or cognitively sophisticated than Whites, we could utilize empirical tests from psychology, sociology, and cognitive sciences, in order to establish the racist's initial moral judgments to be noncredible given their basis on false propositions.

included as theory filters in the filtration procedure, can determine what these facts are. The white supremacist might scramble for some other features, if not intelligence, upon which to base his initial judgment of white superiority. He may be hard-pressed to find a plausible candidate, however.

In this way, a "systematization of prejudices" is not as easy to attain as Brandt's criticism seems to suggest. Imagine, for instance, that we interviewed a thousand racists of varying shades about their moral judgments concerning the moral inferiority of some race. We would expect to find that most of these racists' views were based upon false facts, pseudoscience, morally irrelevant features (that they themselves could acknowledge as irrelevant, when investigated), or a misunderstanding of concepts. Such initial moral judgments based upon mistaken facts, confused concepts, or morally irrelevant features would be pruned from the set of moral judgments that are to be included in an NRE set as considered judgments.

Established sociological and biological theories are to be included in filtration. For instance, established empirical theories have significantly revised the conception of "race," where certain conceptions of race are found to be ungrounded.[67] If the white supremacist has a naïve concept of "race", where he takes race to be a natural kind with clear racial distinctions that can be drawn based on phenotypic expression – namely pigmentation of the skin, then it seems that if empirically substantiated theories that eliminate our folk concept of "race," the white supremacist will have to revise his view accordingly. No longer is there the concept of "race," at least as he has been conceiving of it, which his moral judgment has been based upon. The racist can revise his concept of race into a more sophisticated and less problematic version, and regenerate his moral judgment, but this moral judgment will be different from his original one, and must also cohere with concepts, facts, and morally relevant features.

If the filtration process is fortified in this way by including scientific theory filters that filter out mistaken facts, conceptual confusions, and morally irrelevant features, this

[67] For instance, longstanding scientific evidence establishes that genetic differences among individuals irrespective of racial or ethnic group membership are greater than differences that exist among individuals taken as representatives of such racial or ethnic groups (Lewontin, 1972). This would lead the filtration procedure to reject an essentialist conception of race, as it doesn't map at all well with genetic similarities and differences that the typical racist presumes.

will typically exclude the initial moral judgments the racist will have. In this way, it would appear that even the method of narrow reflective equilibrium has greater capabilities than critics, such as Brandt, have realized or acknowledged.

The method of narrow reflective equilibrium still might allow for some cases of racist NRE sets, where racist initial moral judgments pass filtration to become considered judgments. If so, the racist NRE set could be on par with a non-racist NRE sets. We might imagine, for instance, that a non-racist has CJs about the equal status of non-whites, which match her principles, and vice versa, and yet her reflective equilibrium is just as coherent, and thereby just as "justified" as the white supremacist. Though we would like to grant the non-racist's NRE set greater credibility than the racist's NRE set, Brandt would assert MNRE has no resources to make this distinction. If this assertion proves true– even in rare cases such as in a case of a "sophisticated racist" whose moral judgments pass filtration – there would seem to be a problematic limitation with relying upon mere two-tiered coherence between CJs and moral principles.

A sophisticated racist, for example, might have a relatively informed view of scientific facts and concepts of race. He might eschew the typical essentialist view of most racists, deeming such a biological "natural kind" as naïve, and take a more nuanced and sophisticated view, basing the category of race upon populations, for instance, and acknowledging that race categories are overlapping and imperfect. He might base his views upon some sophisticated scientific evidence that, while controversial, provides some empirical basis that has not been entirely discounted. In such a case, the sophisticated racist's views are not based upon clearly evident factual mistakes and do not seem based upon a lack of conceptual clarity (though the concepts themselves might be, for the time, unclear). Given that the sophisticated racist seems to satisfy relevant cognitive conditions when generating his moral judgment, we cannot entirely discount the sophisticated racist's NRE set, and so Brandt's criticism that MNRE represents a "systematization of prejudices" applies. According to the method of narrow reflective equilibrium, there could be no arbitration between the sophisticated racist's NRE set and the non-racist's NRE set until the facts and concepts they were in disagreement about were settled, and one of their initial judgment sets, regarding race, would be discarded

due to new substantiated theory filters inducted into the filtration procedure. Until then, both NRE sets would enjoy equal status as validated in reflective equilibrium.

Brandt's criticism, even if only partially successful, shows the necessity of including a third set into MRE: namely, background theories. Background theories are, in part, the battlefield where competing RE sets gain or lose ground in regard to coherence. Background theories include scientific theories that are not decisively substantiated enough to serve as theory filters in the filtration process, since theory filters must be acceptable to all rational and reasonable individuals who are moral subscribers.

In relation to our case of the sophisticated racist versus the non-racist, background theories would help adjudicate between the two competing NRE sets. Generally, if there are two NRE sets that are competing, and one is based upon scientific theory A, while the other is based upon scientific theory B, then the degree of validity each respective NRE set enjoys depends on the degree of substantiation of the background theory that coheres with the NRE set, and is inversely proportional to the degree of substantiation of the background theory that conflicts with the NRE set. For instance, in regard to scientific facts, the NRE set of the sophisticated racist who bases his view on the science included in books such as *The Bell Curve* (1994) would have his view fortified to the degree the book's scientific hypotheses turn out to be corroborated, and against the degree alternative and contrary hypotheses turn out to be corroborated. This also applies to concepts. To the degree that a "race" can be legitimately conceptualized as a sufficiently discrete category,[68] the sophisticated racist's NRE set will be corroborated as coherent. Contrarily, to the degree that a "race" is scientifically corroborated as a mere social construct, a grossly ambiguous category, or a group quite different from the category the sophisticated racist is referring to in his moral evaluation, then the sophisticated racist's NRE set is proportionately discredited.

Of course, background theories are not limited to scientific theories, but also include metaphysical theories and normative theories as well. I will discuss the nature and function of background theories further in the following section.

[68] For instance, by providing some multi-locus of genetic similarity of functional parts of DNA between phenotypically similar individuals.

Background Theories and MWRE

Supplementing MNRE, comprised of the doubled set – considered judgments and principles – MWRE includes a third set: background theories. The introduction of this third set initially provides more "traction" to MWRE. Thereby, MWRE can more effectively avoid charges of trivial circularity or arbitrariness.

Background theories can provide further traction for MRE in two main ways: (1) the inclusion of BTs requires a more stringent coherence of reflective equilibrium, given that this coherence must obtain between three distinct sets, rather than two (2) background theories may provide further independent support, stemming from theories in science, social sciences, and metaphysics, as well as normative theories.

The inclusion of background theories in MRE is a way to further ensure that moral principles are not mere accidental generalizations of considered judgments. This is similar to scientific methodology, which aims to distinguish mere accidental systemization of data from scientific laws. Daniels explains, "In science, we have evidence that we are not dealing with accidental generalizations if we can derive the purported laws from a body of interconnected theories, provided these theories reach, in a diverse and interesting way, beyond the "facts" that the principle generalizes" (1996a, pp. 82-83) Similarly, principles in MWRE can ascend the status of mere accidental generalizations if the principles cohere, rather than conflict, with a body of substantiated background theories.

Rawls presents several examples of background theories: a theory of a person, a theory of procedural justice, a theory of the role of morality in society, the ideal of a well-ordered society, and general social/psychological theory. A more extensive conception of background theories is not readily extractable from Rawls' work, mainly because Rawls' conception of background theories is specified to his interest in explicating his system of justice as fairness.

Rawls' proposed theories fall into three basic categories: normative theories (e.g., the role of morality in society), metaphysical theories (e.g., theory of a person), and social

scientific theories (e.g., psychology).[69] Some of these BTs are moral (e.g., theory of procedural justice) and some are nonmoral (e.g., social/psychological theory).

From this store of initial background theories, we can extrapolate other theories to include as background theories. Though Rawls omits mentioning scientific theories, there's no principled reason that scientific theories would be excluded; if a theory bears upon an NRE set, it can serve as a background theory. There are only two general criteria that background theories need to satisfy in order to be included in WRE: (1) that the theories are relevant to the acceptability of CJs or sets of principles, and (2) that these theories are in some way independently substantiated.[70]

Expanding MRE to include background theories provides further demands for coherence: the main criterion for appealing to a BT in MRE is to determine whether or not the BT coheres with an NRE set. If the BT coheres with an NRE set, the BT gives more traction to the NRE set in question by providing itself as a third set, as well as by providing further independent, possibly objective, grounding. If the BT does *not* cohere, then the NRE sets are subject to revision. This incorporation of a third set, a background theory, cannot be merely a trivial inclusion, which would not assist in adjudication. A BT would be "trivial" if it was merely a restatement of the moral principles.[71] Also, "theories" which are not supported in any way, would not provide any traction.[72]

One upshot of the demand for coherence in the tripled set of WRE, is that MWRE provides additional resources to respond to the objection that two people can have two equally coherent, but quite disparate, NRE sets, with no way to adjudicate between the two. First, the introduction of background theories into reflective equilibrium means such dilemmas are far less likely to occur, as we would expect one NRE set to be more coherent with BTs than the other. Second, the inclusion of background theories can

[69] Some theories seem to be a moral/non-moral hybrid. The role of morality in society, for example, seems to relate to both the how morality *should* be included in society, and in what ways it's feasible to incorporate it in society.

[70] In illustration of the satisfaction of these two criteria, I will examine Rawls' consideration of the theory of the person in relation to utilitarianism and Kantian deontology, later in this section.

[71] Another way a BT might be trivial is if the BT cohered with any possible NRE set; in such a case, there would be little traction provided by its conclusion, as it would be universally coherent. One way an NRE set is corroborated is via coherence with a substantiated BT that is *not* coherent with other NRE sets.

[72] We might imagine some metaphysical theories being coherent with an NRE, but being completely unsubstantiated as "theories," even in Daniel's self-acknowledged loose sense of the term.

significantly contribute by providing further independent, possibly objective (and sometimes empirical) grounds; this provides further means by which to adjudicate between competing NRE sets to determine which NRE set is more substantiated via a widened coherence.

For example, a Kantian and a utilitarian may both have equally coherent NRE sets. By appealing to background theories, it might be possible to vindicate the Kantian NRE set and discredit the utilitarian NRE set. Rawls discredits utilitarianism on this basis, asserting that utilitarianism doesn't cohere with a theory of persons, which regards our commitments regarding the metaphysical boundaries of persons: "…there is a sense in which classical utilitarianism fails to take seriously the distinction between persons" (1971, p. 163) The theory of persons is a metaphysical background theory; the metaphysics of this theory is partially dependent upon other scientific background theories, such as biology.[73] In this way, whether the Kantian or the utilitarian position is the most coherent in reflective equilibrium depends, in part, on a background theory.[74] For instance, if Derek Parfit's argument (1987) proves better substantiated on the issue of metaphysical personal boundaries, then that argument would provide a gain for utilitarianism (at the expense of Rawls' Kantian conception of the person).[75] The point, here, is not to engage, much less resolve, this dispute; the point is merely to show that such disputes are, at least in principle, tractable, given the inclusion of background theories in MWRE.

---

[73] Some metaphysical theories are better than others: for example, if a metaphysical theory is more internally coherent than another, or doesn't presume or result in absurdities or paradoxes. In relation to the theory of persons, we can imagine the "Theseus's Ship" problem applied to the human body and its identity, which is at the intersection of metaphysics and biology. If the material constituting our bodies can be said to have been entirely replaced by the end of every seven years, then can we be said to be the same person? What if the old cells are somehow sustained and constructed into a "replica" at the end of the seven years? Such considerations – as well as additional considerations regarding how we process temporal experience, memory, and so forth – do motivate some metaphysicians, such as Parfit, to abandon certain metaphysical conceptions of the person, such as a possibly "naïve" Cartesian paradigm, for less intuitive but more internally consistent and widely coherent metaphysical positions.

[74] The theory of a person is a metaphysical theory that is informed by empirical evidence; therefore, it can be considered an, at least, partially objective background theory which satisfies Daniel's independence constraint.

[75] If Parfit's conception of the person, promoting some version of a "bundle-theory" of identity, decisively decided the issue, this metaphysical theory would no longer count as just a background theory, but would also be included as theory filter in the filtration process, weeding out initial moral judgments that were based on a different conception of the person.

In cases where no present background theories could adjudicate between two disparate but equally coherent WRE sets, this possibility needn't impugn MWRE as a moral methodology. Scientific methodology countenances the coexistence of two competing models, both of which accommodate the data with similar success and conform to scientific background theories with comparable coherence. The fact that there are two models doesn't indicate that the methodology is invalid. The coexistence could indicate either that both theories are, so far, equally substantiated models[76], or it could simply indicate that the body of evidence so far is indeterminate, in which case we have to wait for further background theories or new data to bear out which hypothesis coheres better.[77] If we are to presume there is one determinate world, we might expect background theories to eventually provide facts that would corroborate one WRE as superior in coherence to the other. This expectation might be too optimistic in a world where there is underdetermination of any narrow or wide set: we might never arrive at a single coherent moral theory through MWRE. If this occurs, this might be reason for us to be pluralistic in some of our moral principles; this might be likened to scientists who are epistemically pluralistic in scientific hypotheses, believing that even if there is one determinate world, empirical evidence may underdetermine which scientific hypotheses are superior to others.[78] In such cases, it seems sensible to be, in the very least, epistemic pluralists in science as well as moral philosophy. Of course, this is a stance that one needn't have at the outset, but is a stance one would reasonably adopt if it turned out that repeated applications of MWRE resulted in equally coherent but disjoint WRE sets.

---

[76] In which case, we might assert ethical pluralism on a metaphysical or epistemic basis. We could assert, similar to the arguments posited by some philosophers of science in relation to science, such as Ronald Giere (1999), that a tripled set of principles, CJs, and BTs, are just one model among other valid models, and that given our epistemic limitations, this is the best we can do. I, however, would argue for the second option: that is that there is one determinate moral world, and though theories may occasionally compete, empirical evidence or wide coherence will eventually favor one over the other.

[77] Newtonian mechanics was competing with Einstein's generally relativity for a while, but for the matter to be settled, further empirical evidence had to be introduced with was not readily available. After the Eddington experiment, previously mentioned, the measurements of light curvature around the sun was able to corroborate Einsteinian theory while discrediting Newtonian theory.

[78] We might consider the ostensible contradiction of Newtonian mechanics and quantum mechanics: the former where causal laws dominate, and the latter where causal laws are rejected for probabilistic regularities.

Presuming our notion of non-relativistic moral objectivity, we might expect MWRE to arrive at a singular most coherent WRE set.[79]

Moral Background Theories

In order to achieve traction in MWRE, background theories must have some degree of independent support. Daniels (1979a, p. 259) calls this the "independence constraint," explaining: "Some interesting, nontrivial portions of the set of considered moral judgments that constrains the background theories and of the set that constrains the moral principles should be disjoint." In regard to scientific and other nonmoral background theories, the independence constraint is easy to satisfy. In regard to moral background theories, however, this independence constraint is more difficult to satisfy.

A moral principle can only function as a moral background theory, as long that moral BT satisfies Daniel's independence constraint, where the set of CJs that support it as a BT is significantly disjoint from the set of CJs that support the moral principle under consideration. Expectedly, the degree of independent substantiation of the moral background theory will vary proportionately with the degree of disjointedness between the two CJ sets: the set of CJs that systematize to the moral background theory and moral principle, respectively. The substantiation of the moral background theory will also be a function of the number of considered judgments it coheres and conflicts with – not only of a singular individual, but those of others as well.

In illustration, consider an example where a version of the harm principle is serving as a moral background theory.[80] Imagine the principle represented in the prescription, "Do not cause undue harm to others." This moral BT can check a moral

[79] In a similar way, many scientists expect the progression of science to whittle down the number of scientific theories into fewer and fewer theories that are broader in scope. Some scientists dream of a grand theory that unifies all disjoint theories. Certainly in the history of science, such unification has occurred in several areas.

[80] Another moral principle that could serve as a background theory might be the principle that asserts that individuals ought to be treated in accordance with their deserts, where innocent individuals (such as animals) deserve no maltreatment.

principle, such as one that asserts, “Consuming factory-farmed meat is morally permissible.” If a meat-eater is investigating this latter moral principle, systematized from various considered judgments that meat-eating seems ok, then this individual may observe a conflict between his moral principle, permitting meat-eating, and the harm principle as a moral background theory (systematized from his CJs, though a significantly disjoint set from those systematizing the harm principle). In the instance of such a conflict between a moral principle and a moral background theory, it appears that some revision is necessary of the individual’s WRE set.

Consider a case of coherence. Imagine Michael has the moral principle that individuals should not be used as a mere means to an end, which is a principle supported by a set of CJs.[81] In addition, Michael has the moral principle that individuals have natural rights, which is supported by a different set of CJs. Though there may be, perhaps, some overlap between these sets, nevertheless as long as the two MPs satisfy the independence constraint, the means/end MP is further substantiated via coherence to the moral background theory concerning natural rights; in addition, the moral BT of natural rights seems likewise corroborated via this coherence.

As seen above, multiple moral principles can cohere or conflict with one another, where one of the moral principles functions as a moral background theory. In cases where the first MP is constrained by a completely disjoint CJ set than the CJ set that constrains the second MP, any resultant coherence seems to provide some substitutive support to each of the two moral principles. Moreover, this coherence is stronger than the coherence that would obtain between two moral principles that, though sufficiently satisfying the independence constraint, shared some considerable overlap of supportive considered judgments.

Normative Background Theories

[81] And these considered judgments that constrain this moral principle are not CJs related to “rights” considerations; instead, for instance, the CJs might stem from some empathic cognizance of desiring not to be treated that way oneself.

Some background theories are a mixture of empirical and moral theories. I will refer to these BTs as normative background theories. Daniels (1979a, p. 261) cites Rawls' theories of the person and of the role of morality in society as two background theories that properly satisfy the independence constraint, as the normative portions of these background theories are not derived from the same set of CJs that abstract to or constrain the moral principles under consideration.[82]

In explication of normative background theories, consider the BT of the role of morality in society. This normative background theory is a theory based upon at least two considerations: (1) What role morality could have in society. (2) Given what role morality could have in society, what role *ought* morality to have in society? In this way, the role of morality in society seems to be a background theory constituted by both nonmoral and moral theory: psychological theories, sociological theories, political theories, and moral principles. These theories are inextricably fused into one background theory as it does not seem possible to sensibly answer what role morality should have in society without first taking account of the psychological, sociological, and political facts, as well as the reality of the human condition that is shaped by all of these factors.

Insofar as a normative background theory has one or more moral principles in its content, these constituent moral principles need to satisfy the independence constraint as previously described. For instance, if we are regarding an act utilitarian principle of extreme altruism – e.g., try to maximize utility overall in each action you take – we cannot use a normative background theory that has the same set of considered judgments to support its moral content. We couldn't use, for instance, a theory of a person that suggested human beings should be deemed interchangeable, when that theory of a person (though partially comprised of certain metaphysical theory on personal identity) was based upon utilitarian-related considered judgments (which turned out to be the same set from where the extreme altruism principle was systematized). In such a case, the moral principle and the moral content of the normative BT would not be sufficiently disjoint to provide much coherence in any substantive way.

[82] More specifically, Daniels explains, these background theories rest on no considered judgments regarding rights and entitlements; these background theories then can thereby provide *independent* support to such considered judgments.

To extend this example, this act utilitarianism principle of extreme altruism could be checked by a normative background theory, where the moral content of that normative BT did in fact satisfy the independence constraint. Consider again the normative BT of the role of morality in society. Our CJs of the role morality ought to have in society, given feasibility concerns, would likely include some self-regard, lest we give up everything to help the most needy, and thereby undercut possibility of long-term community sustainability. We could consider what the long-term effects extreme altruism would have on social cohesion and sustainability, as well as if such extreme altruism would even be possible by human beings in particular societal and political contexts. The moral content of this normative principle might also include duties of relationship, reciprocation and cooperative responsibilities. Finding that our normative BT is supported by considered judgments of quite a different sort than the considered judgments of the moral principle (which have an act utilitarian flavor), this independence constraint would be satisfied. However, it would appear that the moral principle of altruism and the normative BT of the role in morality in society conflict, and so one of the two would require adjustment in MWRE.

Nonmoral Background Theories

Nonmoral background theories readily provide potential independent support in WRE, as they are not based on any CJs, and thereby automatically satisfy Daniels' independence constraint. Nonmoral BTs also provide greater objectivity to the degree the independent support they provide is not constituted by moral intuitions, but is constituted by empirical facts or substantiated metaphysics. If the nonmoral BT is scientific, its grounds lie in empirical data. If the nonmoral BT is metaphysical, its grounds lie in non-normative ontology, an ontology often substantiated by scientific background theories.[83]

[83] For instance, we might have a non-substance-dualism ontology given the conservation of energy; determinism given the determinate laws of (non-quantum) physics; or a particular stand on the theory of persons given cognitive science.

Like moral background theories, these nonmoral BTs can adjudicate between competing NRE sets through MWRE. One point to be clarified is how a non-moral BT can bear upon sets of moral judgments and principles. The fact-value gap asserts moral prescriptions cannot be validly derived from nonmoral descriptions.[84] How then could a nonmoral BT bear upon a moral principle?[85]

Moral principles imply facts; if nonmoral background theories change these facts, then this affects the acceptability of principles.[86] To elucidate this further, we can consider a few examples in which facts provided by nonmoral background theories affect principles, and delimit viable NRE sets.

Consider various moral principles that bear upon the treatment of animals. Views regarding the treatment of animals have changed throughout the centuries. Descartes, as well as Kant, believed there was nothing intrinsically wrong with inflicting pain on animals in order to serve human interests. During the 19th century, many scientists thought it morally permissible to practice vivisection on unanaesthetized cats, dogs, and other animal subjects. The discoveries of science, however, have progressively provided evidence of the mental capacities of certain animals. Primates, for example, have been found to possess a high degree of sentience, with capacities for learning -- some even capable of learning rudimentary forms of sign language.

With these facts in mind, let us imagine we have a set of principles. One principle asserts the following: "Moral considerability, all else being equal, should be in proportion with a being's sentience." A second principle is a manifestation of the first, and relates to facts: "Since primates don't have much sentience, it's morally permissible to use primates to serve our minor interests" – such as for cosmetic testing. The first moral principle is an evaluative principle that bears directly upon facts: moral

[84] It should be noted that the arguments that appear to support a fact-value gap can and have been called into question. Nevertheless, some arguments from facts to values clearly are invalid arguments; thus, one would want to investigate the nature of any such argument to see whether or not it was a reliable one.

[85] Margaret Holmgren explores this difficulty (1989, pp. 50-56). I will consider her discussion toward the end of this chapter.

[86] One example of a moral principle that implies certain facts is the qualifier that "ought implies can." If background theories show that an agent cannot when previously we presumed the agent could, then the prescription that the agent ought to do X (when X is not a possible action the agent could perform) is no longer the case: that is, it is no longer prescriptive where it would be considered immoral if the agent does not do X.

considerability in relation to certain facts of subjects. Once the facts are plugged into the function, the resulting moral considerability is determined. The second principle is derivative from the first principle: it is a manifestation of the first principle once "primates" is plugged in. Given this fact-referent principle in our NRE sets, if a nonmoral BT changes what we should say the facts are (say, through scientific discovery), then a principle, in this case the second principle, will be modified accordingly. If we were to submit such a revised view of the facts (which are the fruits not just of one background theory, but more likely of several), it seems that this would and should change our moral conception in a very clear, non-arbitrary way. The individual, seeing his moral principle in jeopardy, might supplant the first principle with some other principle of the form, "Moral considerability, all else being equal, should be in proportion with a being's X," where X describes some morally relevant feature of the subject's capacity. Singer considers various options for X in his work on animal liberation, and argues that features – such as language-use, human DNA, intelligence, and so forth – are unacceptable as morally relevant features as they result in absurdities.[87] The possibilities for X, however, are very limited, and are subjected to rigorous tests relative to our considered judgments. Singer's argument seems to be, at least implicitly, that given the resultant absurdities – the conflicts with our considered judgments – that follow from defining X as anything else other than sentience, other moral substitutions for X will not cohere as well in MWRE. The point here is not to prove Singer's arguments against all alternate MRF possibilities; I wish merely to show that these arguments could be effective, and that Singer's argumentation accords with the method of wide reflective equilibrium.

In further illustration of BT adjudication, let us consider a case where the nonmoral background theory is metaphysical rather than scientific in nature. For the sake of argument, suppose we take determinism as a nonmoral, metaphysical background theory. Determinism could be introduced as a nonmoral BT to adjudicate between two

[87] Babies don't use language as well as chimps; we don't think more genius-level intelligent human beings are morally deserving of more intrinsic moral consideration that "normal"-level intelligent human beings; permanently vegetative human beings have human DNA but we do not find them innately morally considerable as they have no interests; etc.

equally coherent NRE sets: one retributivist-based, the other consequentialist-based. We can imagine how there can be clear, non-arbitrary interplay between this BT and two RE sets. The retributivist RE set might include a set of principles: "individuals deserve to be proportionately punished if they freely decide to commit immoral acts" and "people who freely choose to murder should be put to death by the state." The utilitarian RE set, in contrast, denies any notion of retributive punishment. If determinism were established as a substantive metaphysical theory (and compatiblism and free will theory found unintelligible), then this nonmoral BT would bear upon the aforementioned moral principle of the retributivist: namely, by invalidating a body of its principles, as there is no situation in which the retributive principle manifests: no one commits immoral acts freely, and so punishment based on intrinsic "desert" of this sort is never warranted.[88] Therefore, the retributivist (given determinism as the BT) would have reason to reconsider his moral principles (and likewise considered judgments) that are based upon certain factual assumptions: namely, the assumption that free will, in some robust sense, is a possible feature of human action.[89]

If determinism were decisively established as the only sensible metaphysical position on the issue of human action, determinism would be a metaphysical theory that could be included as a theory filter in filtration, rather than as a background theory in MWRE more generally. If such determinism became so established, we should then resubmit our CJs into the filtration procedure, seeking coherence with this new theory filter. Generally speaking, to avoid error in our moral judgments, we should resubmit our

[88] Though we might have good reasons to pretend to the contrary, given other factors, like human psychology.

[89] I do not intend to pretend that the free will debate is in any sense settled, or even leaning toward one metaphysical camp over another: rather, I merely wish to illustrate how a metaphysical background theory can adjudicate between two competing NRE, invalidating principles and CJs of one, while not necessarily the other. Science might enter into the metaphysical picture, here, by substantiating how plausible it is that determinism is true – or false, in the case of quantum physics. Even if we are not strict metaphysical determinists, the fact that certain environmental factors can significantly influence an agent's desires does carry moral weight. For instance, if the "Twinkie defense" were true (a legal defense strategy that argues for a defendant's diminished capacity due to excessive junk food consumption), it might ameliorate agential responsibility for murder in that case. More plausibly, however, are cases in which brain tumors have affected the "inhibition control" parts of the brain from functioning in certain individuals who then uncharacteristically commit a crime. It would seem this would bear upon our assessment of their moral responsibility.

CJs to the filtration procedure every time a new substantiated and filtration-relevant theory is introduced into the filtration process. In this case regarding determinism, if an individual experienced the common initial moral judgment that a murderer was *intrinsically* deserving of punishment, hard determinism would supply the metaphysical facts that this person's action was not robustly free. Cognizance of this fact would seem to preclude this moral judgment from passing filtration and reaching or maintaining status as a considered judgment: after all, ought implies can, and intrinsic desert – that an agent should be punished because he did X but ought not to have done X – presumes the murderer had a robust choice in the matter whereas he did not.

In farfetched example, we might discover that the Hell's Angels are actually conscious-less robots, programmed by an evil scientist. Whenever a Hell's Angel motorcycles over my rose bushes, I might get angry and shake my fist at him as he rides off cackling. My initial moral judgment will surely reflexively be morally condemnatory, morally evaluating that such hellions intrinsically deserve punishment. *Consideration* of the facts – namely that they are automatons rather than rebels – precludes this initial judgment from passing filtration, however, and becoming a considered judgment. It would be irrational for me to assert "Snake should be punished for trampling my rose bushes!" My condemnation should be redirected toward the evil scientists or a society that allows such robots to live free or die (even if not robustly free).

Traction for Change

Whereas the coherence of one's NRE set might indicate no reason to change one's CJs or principles, inclusion of background theories can provide such impetus. Background theories can provide the needed traction for one to *change* one's RE set.

Take for example the issue of animal liberty. Currently most Americans believe that having animals killed in order supply meat to consumers is morally permissible. A carnivore's CJs might match his principles in his NRE sets. His intuition toward animals would likely extend to prohibition of wanton cruelty, though not of useful exploitation. He might adhere to Kant's 2nd formulation of the categorical imperative, which prohibits

using rational beings merely as means to an end; however, he might fail to recognize animals to be rational to any significant degree.

In this way, the carnivore would have a coherent NRE set, yet by expanding his NRE set to become a WRE set by including background theories – particularly those that bear upon the rationality of animals – might induce the revision of his principles. If nonmoral BTs indicate that animals are more rational than he initially granted, then Kant's 2nd formulation might apply to them.

Animal liberty activists appeal to an assortment of background theories when trying to alter people's moral perceptions. Peter Singer (1975) invokes neurology and cognitive science, citing physiological evidence that animals feel pain in quite the same way human beings do. Singer invokes American history – a history riddled with status quo racism and sexism – as a background theory indicative of our sociological and psychological tendencies as humans, individually or as a society, to be prone toward rationalization concerning exploitive maltreatment of a disempowered underclass. Other animal liberty activists have stressed that our common evolutionary ancestry to animals should shift our moral perception away from anthropocentrism.[90]

In arguing for animal liberty to the unconverted, activists introduce such background theories, which cause uncomfortable tension to the meat-eater's WRE set. These theories provide pressures to individuals, on pain of irrationality, to revise their WRE set in order to retain equilibrium.[91]

Two Examples of Adjudication

[90] Certain psychological theories might be brought to bear as well: for example, perhaps studies similar to the Milgram Experiment might indicate the extent to which human beings can be morally callous to the infliction of suffering upon others – just so long as there is an authority, whether it be a man in a white lab-coat or an entire society of meat-eaters, meat-sellers, and advertisers -- to assuage their conscience.

[91] One, however, needn't necessarily appeal to background theories to promote change in an agent's RE set. For his work on animal rights, Tom Regan often focuses on the stark inconsistency between our attitudes and perceptions towards pets versus towards meat-industry animals. This inconsistency seems to be internal to the principles one has in his/her NRE set.

Background theories can adjudicate between two competing, equally coherent NRE sets.[92] Consider the ostensibly intractable issue of abortion. Imagine two competing, equally coherent NRE sets: pro-life and pro-choice. The pro-life advocate bases her position upon several principles: one is "destroying an innocent human soul, without sufficient extenuating circumstances, is morally wrong." She actually already has a background theory in place, religious in nature, which factually manifests in this evaluative principle. A change in this religious (metaphysical) background theory would invalidate the facts present in this principle, and thereby nullify the principle. For instance, it might happen that while locked in analytic discussion with a godless philosopher, and examining such theories as determinism, the argument from evil, arguments against dualism, divine command theory, and so forth, she becomes convinced of the non-existence of God and the non-existence of the soul, at least as she previously understood the soul. In this case, the introduction of nonmoral background theories, metaphysical and scientific, could dissolve her previous background theory, which was religious in nature. The dissolution of her original BT would – or at least could – reasonably result in her changing her presumption in the status of the fetus as a "human life." This, in turn, would invalidate the way she conceives of abortion: no longer does it involve destroying a sacrosanct "soul." The pro-life advocate might still try to save her principle, that abortions are always wrong, by predicating the principle upon other features, such as potentiality of the fetus. This move might work, but she would have to

[92] The notion of coherence, in scientific methodology and in MRE, is both qualitative and quantitative. To illustrate the distinction, suppose Simon has only two considered judgments, and these CJs perfectly cohere with his one moral principle. This would count as perfect *qualitative* coherence. Quantitative coherence, however, is also to be included within this concept of coherence, where the greater the number of entities present that can cohere or conflict is important. In contrast to the case above, consider a case where Cory has 801 CJs, and four moral principles. Of these 801 CJs, 800 CJs cohere with the four moral principles, and 1 CJ conflicts. In such a case, we would have less than perfect, 100%, coherence, and yet this less-than-perfect coherence would be more impressive than Simon's coherence in the first case. Moreover, Cory's NRE set enjoys greater coherence overall, even though it has less qualitative coherence than Simon's. Thereby, if the two were in competition with one another, Cory's NRE set is superior to Simon's; it is in greater equilibrium. One question that arises is what weights to attach to qualitative coherence and quantitative coherence. Is the quality of coherence more important than the quantity of coherent entities, or vice versa? This is a question for scientific practice as much as moral methodology. Despite a lack of precise specification and weights, what to do seems clear in most cases, when such questions arise.

see if this feature cohered with her other CJs and principles.[93] It should be emphasized that there are a limited number of morally relevant features upon which considered judgments can reasonably be predicated. These morally relevant features are themselves subject to debate: for instance, predication of an anti-abortion CJ upon the fact a human being has human DNA (and all things with human DNA are morally considerable) invites us to test this principle via counterexamples. In this case, we might present the counterexample of an anacephalic fetus who has human DNA but only a brain stem, where, lacking any brain physiology necessary for consciousness, does not seem morally considerable (at least not to the same degree as the mother herself).[94]

Irrespective of the controversy surrounding these nonmoral background theories, the point remains: nonmoral background theories can be relevant in arbitrating between two NRE sets. If the problem of evil was incontrovertible, for instance, the original pro-life NRE set might result in diminished coherence in the face of a subsequently revised set of BTs.

Revisiting an earlier example of this chapter, consider the intractable conflict between the reflective equilibrium sets of the white supremacist and the non-racist. If we just consider the two respective NRE sets, it may turn out that both are equally coherent. However, we can find a way to adjudicate between the two by incorporating a background theory into our reflective equilibrium. For example, the white supremacist's NRE set will need to be reconciled with certain relevant background theories: cognitive science, certain sociological theories, and so forth. Cognitive science research would indicate that the white supremacist's presumption of racial inferiority, say in mental

[93] Judith Jarvis Thomson (1971) considers the potentiality argument, and generates intuitions via thought-experiments to show that our moral intuitions do not cohere with potentiality considerations. This is similar to the previous Singer discussion about candidates for "moral considerability" instead of sentience.

[94] In this example, we are relying upon our moral intuitions to determine whether or not the anacephalic fetus, and any DNA-bearing entity, should be considered morally considerable. I argue that this determination of moral relevance is determined upon the basis of elemental intuitions, concerning what is and what is not morally relevant, that we agree upon with significant overlapping consensus – especially in certain cases. One interpretation might liken morally relevant features to features of grammar, where some word-composites just could never count as sentences, just as some things could never count as morally relevant. In chapter one, I argue that there are certain limited bases upon which a moral subscriber can predicate their moral evaluations, and there are other ways that she clearly cannot. The line of reasoning I present is similar to arguments presented by Philippa Foot (1958; 1959).

capacities, of non-white individuals is false. Moreover, sociological theories, based in genetic findings, that provide argument for the dissolution of the very notion of "race" might undercut the very categories the white supremacist is relying upon in formulating principles and considered judgments. In this way, the white supremacist's RE set lacks coherence with certain background theories, whereas there is greater coherence between the non-racist's RE set and these empirically based BTs.

Wide Coherence and Systematization

Imagine Bob and Sue, acknowledging each other as rational and reasonable moral subscribers, pool together their considered judgments, and thereby have the same set of considered judgments. Though they start from the same set of CJs, they will likely systematize these CJs in different ways, maintaining and abstracting different principles in reflective equilibrium with different sets of CJs. Upon what basis is the stronger WRE set to be determined? The following criteria, though not exhaustive or singularly decisive, contribute to a stronger WRE set:

1. The WRE set leaves out less CJs, as anomalous outliers, from systematization than the competing WRE set.

2. The WRE set depends on fewer moral principles that sufficiently explain/cover the CJs.

3. The WRE set coheres with a greater number of background theories.

4. Those BTs, as a group, with which the WRE set coheres, has more significant substantiation than the BT group of the competing WRE set.

5. The WRE set has a fewer instances of conflict with BTs than the competing WRE set.

6. The BT group that conflicts with the WRE set are less substantiated than the BT group that conflicts with the competing WRE set.

7. Generally, the WRE set is not part of a degenerative program, in a Lakatosian sense, where considered judgments are systematized or "saved" in an irrational way. That is, the better WRE set will tend to have a leaner auxiliary belt of protective hypotheses to take the fall for conflicts, as opposed to a fatter auxiliary belt where weakly

substantiated BTs or moral principles are introduced or maintained in order to explain away conflicts.

As delineated above, granted that two NRE sets are equally coherent, their wider coherence, after the introduction of background theories, determines which of the two systematized sets is more substantiated via coherence.

In illustration, imagine that Bob and I agree upon a coextensive set of considered judgments: I take seriously his considered judgments as deliverances of an individual with proper moral faculties, and he presumes the same of me. However, I systematize this set of our considered judgments differently from him: abstracting and maintaining different moral principles to systematize different groups of CJs, and excluding some other CJs entirely from the systematic groupings.[95] Due to this difference in systematization, despite that Bob and I started from a shared set of CJs, we end up with two different systematized WRE sets.[96]

Let us imagine that Bob ends up, after his systematization, with the systematized considered judgment (SCJ) that homeless persons deserve no social assistance (though he allows that it might be nice, or even virtuous, for an individual to help). His SCJ turns out to be maintained, in part, given Bob's subscription to a metaphysics that emphasizes that human action is free in some sense that allows for moral responsibility. I, on the other hand, have a different WRE set; nevertheless, I can understand why Bob has this SCJ, given his systematization. In determining which of our two WRE sets was superior, I could cite, perhaps, that he poorly systematized our shared CJ set in that the SCJ in question conflicts with a number of substantiated background theories: for instance, that many homeless individuals are mentally ill; that there are not jobs available; that upward mobility is empirically less possible than it seems especially when one lacks a home,

[95] Typically, in a case like this, Bob and I would both exclude some, though different CJs, that are leftover outliers to our respective systematizations.

[96] One reason we might have different WRE sets is that we abstracted or maintained different moral principles. Another reason we might have different WRE sets is that we included different BTs into our WRE set, or weighted the included background theories differently (though weight determination of a nonmoral BT primarily depends on agent-independent substantiation). Ideally, we each should include every substantiated background theory that bears upon CJ and principle revision and adjustment. A WRE set can be rightly criticized if it excludes relevant, substantiated background theories that conflict (or cohere) with it, as this affects its coherence.

address, clothes, and skills; that "learned helplessness" is a psychological feature of human beings; etc. Relating such empirical findings of various background theories, I could show how Bob's systematized considered judgment – that homeless people deserve no assistance – ought to be revised in deference to these substantiated background theories.[97]

Conflictive Systematized Considered Judgments and Degenerative Programs

Bob might listen to my criticism of his systematization, but rather than revising his systematized set to exclude this CJ, Bob might dismiss these sociological and psychological assertions as "liberally biased," thereby fattening his "protective belt" of auxiliary hypotheses with the introduction of this background theory of liberal bias in the sciences. I may rightly increase my criticism of Bob's WRE set, as he has introduced a weakly-substantiated background theory into his BT set, and has dismissed others that seem substantiated, such as research in psychology and sociology. In order to save his SCJ, Bob might even alter his metaphysical view of action, assigning further moral responsibility to human action than seems philosophically substantiated.[98]

My WRE set becomes more corroborated if I can point out that those SCJs of Bob's that conflict with my WRE set turn out to be artifacts of unreasonable systematization, or that his systematization is increasingly representative of a degenerative program, in a sense analogous to Lakatos concerning scientific programs. As part of this degenerative program, a conflictive SCJ becomes less credible. This is true even more so if Bob's conflictive SCJ is a "crown jewel" of his degenerative program: the phenomenon/CJ around which he is warping the rest of his WRE artifice.

[97] For clarity, I am presuming the differing NRE sets, of Bob and I, have equal coherence between moral principles and considered judgments, and only background theory coherence is at issue. Moreover, the example presumes that neither Bob nor I have made a mistake in filtration, where CJs were accepted than should have been rejected. If a mistake were made, the differences in CJs would simply be referential to this mistake in process.

[98] For instance, if he became a "free willist" about human action, this position would be less substantiated and more metaphysically troublesome than compatiblism.

If I can indict Bob's conflictive SCJ as a degenerate SCJ, in that it is a considered judgment that Bob ought to discard in rational readjustment of his WRE set, then my WRE set is thereby strengthened by reconciling this conflict. My WRE set is strengthened by the fact it now enjoys greater coherence, freed of this conflict.

I label an SCJ as "degenerate" if it is a systematized considered judgment that, though provisionally credible as a CJ, should have been excluded from the WRE set by a rational individual during WRE set adjustment.[99] If the individual is systematizing his WRE set in an irrational or biased way, where SCJs survive adjustment when they should have been discarded, then these SCJs can be deemed degenerate. For instance, I can point out that if Bob had actually rationally adjusted and revised his WRE set, without unreasonable bias toward a particular SCJ, he would have ditched that SCJ that happens to conflict with my WRE set. If no such bias or systematization mismanagement were made, then his validly systematized CJ that conflicts with my WRE set would represent a true anomaly that should trouble my systematized judgments/principles (as anomalous but corroborated data outliers trouble a scientist when it doesn't fit with his hypothesis). So even though Bob's SCJ conflicts with mine, if I can show how his SCJ is based upon systematization mismanagement (such as a degenerative program where he saves the phenomenon of that SCJ by ignoring or mis-weighting background theories, or by introducing weakly substantiated auxiliary hypotheses/BTs), then I, in essence, strengthen my own WRE set by defending it from possible conflict with that SCJ, given that I can discredit that SCJ as degenerate.

If I can point out Bob's SCJs or systematized moral principles do not cohere well with independently substantiated background theories, then that gives reason to revise his system – even if he does not. If, to turn the tables upon myself, Bob surveys my systematization of our CJ set, and finds that I have an SCJ that does not cohere well with BTs (e.g., coheres with a few weak BTs, but conflicts with several strong BTs), then Bob has given me reason to revise that SCJ. I may, however, act unreasonably, exemplifying

[99] As Lakatos (1970, pp. 138-140) points out, saving a hypothesis is not necessarily irrational. In deference to his point, I would acknowledge that an SCJ does not become degenerate immediately, but becomes increasingly degenerate as the program of which it is a part becomes a degenerate program over the course of systematization where phenomena are deliberately saved without deference to unbiased and rational systematization (as previously defined).

confirmation bias – discounting strong, conflictive BTs, overly subscribing to weak BTs, and seeking out other corroborative BTs to my WRE set – just to save a favored SCJ.

Scientific and Moral Methodologies

In the examination and justification of MWRE as a moral methodology, I have frequently appealed, via analogy, to scientific methodology. It is not my purpose to push this analogy too far, as I am cognizant of the differences between them. Nevertheless, I believe that the methodologies do share similarities, and that consideration and examination of scientific methodology will help elucidate MWRE as a moral methodology, as well as provide a basis of justification for its validity. Another reason for invoking the analogy of scientific methodology is the similarity of intent: both practices are a quest for knowledge about the world through the organization of phenomena.

Daniels appeals to scientific methodology in support of a form of reflective equilibrium. He explains that if we presume scientific realism:

> "Then we would be justified in claiming that certain central methodological features of science, including its coherence and other theory-laden constraints on theory acceptance (e.g., parsimony, simplicity, etc.), are consensus-producing *because* they are *evidential* and lead us to better approximations of the truth. I have been defending the view that coherence constraints in wide equilibrium function very much like those in science" (1979a, p. 279).

As Daniels notes, some scientific experiments are very simple and direct, such as Galileo's observational experiments measuring gravity via rolling balls down inclines. Other experimental hypotheses and tests are much more complex and theory-laden, such as investigating the existence and behavior of quarks. Daniels characterizes moral intuitions to be "data" more in the theory-laden sense of the latter than the direct observational sense of the former.

As Daniels suggests, scientific methodology can be thought of as employing a wide reflective equilibrium itself,[100] consisting of a tripled set: data, hypotheses/theories,

[100] As opposed to foundationalist in methodology, where there are certain fixed and unrevisable Archimedian points upon which to build. According to wide reflective equilibrium methodology, everything is up for revision and, possibly, elimination.

and background theories. Data is generated via observation or experiment, under certain conditions intended to reduce error.[101] A hypothesis is constructed and tested. The data is compared to other data sets, checking for consistency. The hypothesis is examined in reference to other corroborated scientific hypotheses/theories: does the current hypothesis cohere or conflict with other corroborated hypotheses and theories?

For example, consider testing a hypothesis that asserts that the volume of gas at a constant temperature is inversely proportional to the pressure. Imagine our experimental results conflict with this hypothesis. First, we might double-check the conditions of our experiment to determine whether there were any error-disposed conditions that might result in error-disposed data: for instance, we may double-check to ensure the temperature was in fact constant during the experiment. If we find the temperature fluctuated, we would filter out the initial data, deeming it noncredible. Suppose, however, that we found no error-disposed conditions present, and yet the data, resultant from the experiment, still conflicted with our hypothesis. We would then determine how corroborated our hypothesis was with other data, not only of our own repeated experiments, but the results of other scientists as well who ran the same experiment. If we found the hypothesis was highly corroborated from all of the data sets, and yet the present experiment's data was singularly anomalous to this hypothesis, rationality would prescribe we consider this anomalous data is noncredible. We would do this under the presumption that the anomalous data arose under error-disposed conditions, though we were unable to identify the source of error.[102]

We would further find the hypothesis highly corroborated if it cohered with other highly corroborated hypotheses. For instance, the ideal gas law is a highly corroborated hypothesis in that it strongly coheres with experimental data sets, and also coheres with other "background theories," such as kinetic theory. Kinetic theory explains macroscopic properties of gases, such as pressure, temperature, and volume, by considering their

[101] These experimental conditions are set in place before the data is generated in order to reduce initial error. If the data generated under controlled conditions turns out to be anomalous to previous data sets or a corroborated hypothesis, the scientist might look again for error in the experimental conditions. Oftentimes the type of anomaly may suggest where the scientist should look for error-disposed conditions in the experimental setup.

[102] This relates back to Duhem's thesis, as previously mentioned.

molecular composition and motion. Kinetic theory explains pressure in terms of collision between molecules rather than in terms of static repulsion between molecules, which was Isaac Newton's conjecture. In this way, experimental data coheres with the prediction of the ideal gas law, and conforms to kinetic theory as a background theory.

Scientific Methodology and MRE: Shared Criteria

Some of the criteria concerning what constitutes good scientific methodology will carry over to what determines good moral methodology. Characterizing good science, the following criteria are often mentioned: corroboration, falsifiability (or revisability), control, consilience, simplicity, prediction, unification, and fecundity, among others.

We can consider MRE in relation to most of these criteria. Corroboration, for instance, is the primary focus of MRE, as we have already seen. Likewise, falsifiability, or revisability, is present in that all principles and judgments are subject to revision in that they can be falsified and discarded. Control refers to actively seeking data under different conditions, as opposed to mere passive reception of data; this is in order to ensure there is a representative sample of data, to limit sources of error that might arise under similar conditions, and to preemptively combat empirical bias.[103] Consilience is represented by coherence between moral principles or considered judgments, and background theories.[104] Simplicity relates to the principle of parsimony, which we have already considered in relation to Ockham's razor: simplicity prescribes selecting the simpler of two competing hypotheses, presuming both equally cohere with the data, as a way of reducing opportunity for error. Prediction occurs when a moral principle is

[103] Confirmation bias, for instance, is a bias a scientist might have, in which he accepts corroborating data but ignores, either deliberately or obliviously, disconfirming data. This is why scientists are encouraged to try to falsify their hypothesis by subjecting it to rigorous tests. Psychological studies suggest that human beings suffer from confirmation bias in regard to several parts of their lives: moral responsibility, political ideology, personal judgment of others, and so forth. Control, then, is closely related to error filtration in both scientific and moral methodology.

[104] E. O. Wilson (1998, p. 11) writes: "Disciplinary boundaries within the natural sciences are disappearing, to be replaced by shifting hybrid domains in which consilience is implicit. These domains reach across many levels of complexity, from chemical physics and physical chemistry to molecular genetics, chemical ecology, and ecological genetics." This wider coherence occurs in relation to moral methodology as well in terms of background theories: such as the theory of the person in relation to coherence with moral principles such as utilitarianism or Kantian ethics.

corroborated by a considered judgment not included in the set of CJs from which the principle was abstracted. A principle could be said to be predictive when we find considered judgments in other cultures that can either corroborate and discredit the principle, or when we generate considered judgments in response to novel contexts: such as ethics in new applied contexts such as business ethics, medical ethics, and so forth.[105]

Other scientific criteria that may be analogously applicable to our moral methodology are unification and fecundity. Unification is applying a small-set of problem-solving hypotheses to make sense of empirical data (for example, evolution via natural selection to explain different beaks of finches around the Galapagos Islands, similar physiology between animals, and vestigial organs). Philippa Foot's doctrine of positive and negative duties unify our moral intuitions of several varied cases in a much more unifying way than the doctrine of double-effect, which she criticizes as overall conflictive with many reasonable moral intuitions.

Fecundity is when a scientific hypothesis leads to other discoveries or scientific developments: for example, Darwinian evolution has yielded new theoretical disciplines, such as population genetics (Kitcher, 1998a). Moral hypotheses can lead to other discoveries or developments in the field of moral philosophy. Feminist ethics, for instance, has birthed several offspring: theories of race and class; theories of sex identity inclusive of transgendered and hermaphroditic subjects; and even epistemic theories that extend beyond mere value theory, such as experiential knowledge in contrast to propositional knowledge. Even feminist ethics itself has flowered into varying fields, such as feminine ethics, lesbian ethics, and so forth.

A criterion, previously mentioned, is consilience. Over time, valid scientific theories tend to merge and intersect. Geological theories of continental drift merge in coherence with anthropological histories of migration; Mendelian genetics merged with Darwinian theory of species change; psychology seems to be progressively merging with cognitive science, and economics appears to be merging with psychology. In a similar way, consilience seems to appear among moral theories. Rawlsian justice as fairness

[105] A few examples include global division of labor, cloning, genetically modified animals, plastic surgery, etc.

merges with Kantian concepts as well as other moral notions. Rule utilitarianism merges with virtue ethics, in determinations of what sort of life is a valuable human life, which human beings can and ought to pursue. Investigations into certain moral emotions – such as gratitude, forgiveness or contempt – become unified with virtue ethics, concerning what emotions are necessary to cultivate as a habituated character traits, such as self-respect. Lastly, moral theories appear to be progressively merging with nonmoral theories more now, than in the past: experimental moral psychology, economics, cognitive science (e.g., fMRI studies), and evolutionary biology. Gilbert Harman appears to note this trend of consilience:

> "There have been three good trends in moral and political philosophy over the last fifty years or so. First, there has been a trend toward rejecting special foundations, a trend that is exemplified by the widespread adoption of the method John Rawls adopts, in which particular judgments and principles are adjusted to each other in an attempt to reach 'reflective equilibrium.' Second, there have been attempts to use intuitions about particular cases in order to arrive at new and often arcane moral principles like that of double effect, as in discussions of so-called trolley problems. Third, and perhaps most important, there has been increased interaction between scientific and philosophical studies of morality, as for example in philosophical reactions to psychological accounts of moral development and evolutionary explanations of aspects of morality" (2003, p. 415).[106]

The analogy between scientific methodology and MRE proves compelling upon examination. However, we mustn't be oblivious to the differences. One disanalogy MRE has with science is that scientific hypotheses/theories are supposed to be explanatory: they explain *why* the phenomenon exists or behaves the way it does. Moral principles, on the other hand, unify and codify our considered judgments: however, our moral principles are typically not thought to *explain* our considered judgments. Stated another way, it seems that we wouldn't assert that moral principles *cause* considered judgments, though we would readily assert that gravitational forces cause the apple to fall from its tree.

Though this disanalogy seems right, one might argue that this disanalogy isn't as vast as we might originally conceive. A moral philosopher might reasonably suggest that one motivation for moral theory in the first place is to explain why our moral judgments are true. For instance, the doctrine of double effect is proffered as not just an abstraction

[106] My emphasis.

of systematized moral intuitions, but as an explanation of *why* we have such intuitions to start with.[107] In bridging the gap from the other side, a philosopher of science might contend, for instance, that gravity does not *cause* the apple to fall: gravity is merely the generalized description of objects falling. Gravity doesn't thereby *explain* anything: it merely describes a class of things. A moral philosopher, advocating bridging the yawning gap between moral and scientific methodologies, might suggest that moral principles *do* cause considered judgments in the same way that the law of gravity causes apples to fall from their tree.[108] It might be suggested there are laws of moral reality just as there are laws of physical reality: we presume the physical phenomena of the universe are governed by a set of laws, which we can arrive at through observation of the physical phenomena. Likewise, we might presume the moral phenomena of the universe are governed by a set of moral laws/principles, which we can arrive at through observation of the moral phenomena: namely, considered judgments. Through these above considerations – entertaining that scientific and moral methodology might be closer than we initially think – I am not trying to deny significant differences between the two methodologies. I merely wish to point out the possibility that what are commonly seen to be distinct differences between the two might lie rather within our paradigms of perception and presumption than in true differences themselves.

As we have seen, science and MWRE share similarities in methodology. We might take moral intuitions as partial analogs to observational data in science. Daniels (1979a, p. 273) explains in a rather lengthy but significant passage how MWRE is similar to scientific methodology:

> "The accounts of initial credibility we accept for observation reports…are based on inferences from various component sciences constrained by coherence considerations. Observation statements are neither self-warranting nor unrevisable, and our willingness to grant them initial

[107] Mathematical proofs are taken to "explain" why certain theorems are true, even though the causality involved is more similar to Aristotle's formal causality than his efficient causality.

[108] For instance, consider a consequentialist moral principle that asserts "increasing utility is morally good," presuming nothing else morally relevant is at stake (e.g., duties, virtues, etc.). We might take such a moral principle as explaining why we have all these moral intuitions in the first place. Such a principle is more than a mere abstraction of those intuitions, but seems to be an explanation of those intuitions: We have those intuitions because the nature of morality is one where, provided nothing else is at stake, utility should be maximized.

> credibility depends on our acceptance of various other relevant theories and beliefs. Such an account is also owed for some set of moral judgments, but it too will derive from component theories in wide equilibrium. Similarly, in rejecting the view that wide equilibrium merely systematizes a determinate set of moral judgments, and arguing instead for the revisability of these inputs, I suggest wide equilibrium closely resembles scientific practice. Neither in science nor in ethics do we merely 'test' our theories against a predetermined, relatively fixed body of data. Rather, we continually reassess and reevaluate both the plausibility and the relevance of these data against theories we are inclined to accept. The possibility thus arises that these pressures for revision will free considered moral judgments from their vulnerability to many of the *specific* objections about bias and unreliability usually directed against them."[109]

As Daniels notes, just as scientific methodology accounts for initial credibility of observation reports from inferences from various background theories constrained by coherence considerations, MWRE can account for initial credibility in a similar way.

Initial Credibility and Moral Objectivity

Initial judgments are initially credible insofar as they are presumed to be as truth-tracking, where they provide some kind of access to objective moral right and wrongs. Margaret Holmgren (1987, p. 108) makes this point clear, "…if WRE is used as a procedure for formulating an *adequate* moral theory, then it does presuppose the existence of objective moral truths." The existence of objective moral truths, however, cannot be validated within WRE itself; Holmgrem explains, "we must take at least a provisional stand on this issue before we can set out to formulate an adequate moral view" (p. 123). Thereby, moral objectivity is an essential starting assumption of MWRE. In addition, MRE attributes initial credibility to our moral judgments, in that they are presumed to give us some kind of access to moral objectivity.

I contend – as Holmgren does – that these starting assumptions shouldn't be considered too controversial or troubling for MRE, given that we typically do in fact attribute initial credibility to many of our pre-systematized moral intuitions as indicative of some kind of moral objectivity.[110] For instance, the oft-cited intuition "torturing

[109] His emphasis.

[110] The essential assumption of preliminary credibility is just the assumption of initial credibility plus the assumption of the validity of the filtration process. It shouldn't be problematic to focus on preliminary credibility rather than initial credibility as the essential assumption, since the validation of the filtration is a

children for fun is wrong" strikes us as intuitively compelling, as do many other intuitions we have. In fact, most of us would assert that this normative statement is true.[111] In addition, our practice of moral interlocution and argument defers to moral intuitions: for instance, a standard practice of ethicists in argument against moral theories such as utilitarianism and Kantian ethics is to appeal to intuitions that run counter to those theories.

G. E. Moore (1903) characterizes moral intuitions as direct apprehensions of moral features in the world. Historically, intuitionists, such as W. D. Ross (1930), have asserted that moral intuitions are "self-evident," or that they are self-validating normative claims. Fortunately, MRE needn't make such strong claims about the initial status of considered judgments.[112] Considered judgments are merely provisional starting points, still subject to revision. A considered judgment can always be set aside, on the basis that it is inconsistent with a highly corroborated principle.[113] Even if a considered judgment *feels* normatively forceful, it can still be bracketed from the set of considered judgments on the basis of inconsistency with a highly corroborated principle. Considered judgments are never to be entirely rejected after filtration, but can be set aside when they are inconsistent with highly substantiated principles or background theories, or even a mass of other inter-coherent considered judgments.

This is similar to scientific practice in the circumstance where a set of data is inconsistent with a highly corroborated law. If a scientific hypothesis is highly corroborated by sets of data, if there is a data set that is inconsistent, then this data set is typically put aside. Likewise, if a single data set proves to be an outlier far outside the close grouping of all the other data sets, then the outlying data set will be discarded. It is

---

different question than the essential assumption of initial credibility. I argue for the validation of the filtration process in chapter 3.

[111] What I mean by "true" here assumes a correspondence relation between our judgments and the world. Assuming, as previously explained, a version of moral objectivity that carries with it a non-relativistic notion of truth, MRE provides an epistemic methodology that drives toward verisimilitude between our moral judgments/principles and objective normativity.

[112] Daniels asserts that considered judgments in WRE can be given even weaker initial justification than considered judgments in NRE, due to the inclusion of background theories that can provide further support via widened coherence (1979, pp. 259-262).

[113] One reason to bracket a considered judgment is that one suspects an error-disposed condition was present during the intuition's initial generation, but that individual is unable to locate this error-disposed condition.

generally presumed that this anomalous data set generated under unidentified conditions of error: that experimental parameters were not met, that all the relevant experimental conditions were not known or satisfied, that the instruments measuring experimental reactions were not precisely calibrated, and so forth. In order to set aside this anomalous data set, the reason needn't necessarily be known; the mere fact of inconsistency is sufficient to generally presume some error in the experiment that generated the anomalous data set.[114]

In the context of MRE, initial judgments must have some *prima facie* credibility if, after filtration and adjustment in reflective equilibrium, the moral judgment is going to yield a normative result. Initial judgments, then, must provide some kind of indirect access to – or evidence for – objective moral truths. MRE needn't be committed to any particular view of objectivity; MRE must only rely on *some* kind of assumption of moral objectivity.[115]

Relating objectivity in science to objectivity in MRE, Daniels writes:

> "I have been defending the view that coherence constraints in wide equilibrium function very much like those in science. If I am right, this suggests that we may be able to piggy-back a claim about objectivity in ethics onto the analogous claim we are assuming can be made for science." (1979a, p. 279).

Explicating the multitudinous and ambiguous concept of objectivity, Daniels (1979a, p. 274) states: "…given the area of inquiry, claims are thought to be objective if there is some significant degree of intersubjective agreement on them. Second, claims are also said to be objective if they express truths relevant to the area of inquiry."

---

[114] One concern that arises is in terms of falsifiability: after all, if a scientist is discarding all data that conflicts with her hypothesis, how can that hypothesis ever be falsified? I will consider this concern later in the chapter. In regard to a different point, an anomalous data set needn't be necessarily indicative of error. If the hypothesis posited is only asserted as a possible model to approximate outcomes, high corroboration will establish it as a legitimate model for most of the phenomena. The ideal gas law could be considered a reliable model in this way – though only reliable within certain parameters. "Error" changes meaning under this paradigm.

[115] Margaret Holmgren calls this assumption one of "objective moral truth" (1987, pp. 108-125); I am referring to the assumption as moral objectivity, as it seems to tempt the misunderstanding that this assumption must include moral realism.

As Daniels' suggests, the provision of initial credibility can be even weaker in MWRE, than in MNRE. By including background theories that confer independent, possibly objective support, CJs can become substantiated through further coherence.[116]

Broad Systematized Considered Judgments and Credibility

An additional source of credibility, which can occur at any point after filtration, is the inclusion of others' CJs, SCJs, moral principles, and background theories.[117]

For example, systematized considered judgments are those CJs that are systematized within the individual's reflective equilibrium set, whether narrow or wide. Systematized considered judgments become *broad* systematized judgments when they are found to be also coherent with the RE sets of others. This is similar to scientific methodology: if coherent empirical data sets are generated in different experiments in different laboratories across the world, in different countries, by different individuals, under slightly different experimental conditions perhaps with slightly different parameters, this would add credibility to the original data set via broadened coherence.[118]

An individual's SCJ has not been systematized by *other* individual's moral principles and background theories.[119] However, if the two individuals discover that after adjustment in RE that they share the same systematized considered judgments, then this broad coherence further substantiates the credibility of each of their SCJs, as well as their moral principles. This would be especially salient if the other individual had different moral principles or background theories, and yet still arrived at a highly similar set of systematized considered judgments. Even if the other individual has identical CJs, MPs,

---

[116] One clear way coherence provides further credibility is as an indicator: it indicates freedom from error that could result from cultural, historical, ethnocentric, and even biologically-based bias. This presumes, of course, the background theories are free from such error and artifacts; in this way, objective (scientific or metaphysical) nonmoral BTs would be more readily evidential.

[117] Moral principles, after adjustment in MRE, should also be attributed increased "systematized" credibility in that they have survived adjustment via coherence.

[118] One worry that might arise is that this broad coherence is just a result of pervasive error: such as a psychological bias that human beings might be said to have. While a valid concern, the broadened coherence provides another resource for weeding out error, even if it doesn't guarantee certitude, much less objectivity.

[119] Though they could, in which case they might be considered broad considered judgments, though those CJs have not as of yet been systematized.

and BTs, this would still lend additional credibility due to the fact that error is more likely to be weeded out, such as some unidentified individual bias in adjustment between CJs and MPs.

It's true that the same CJ hasn't been systematized by the *other* individual's BTs and MPs, but if they had identical CJs which were so systematized with different BTs and MPs than the original individual, then these differences would only serve to further substantiate the credibility of the original individual's SCJs. Even if the other person's RE set was almost identical to the original person, the fact that the adjustment between CJs, MPs, and BTs, though rational, is still somewhat undefined, the resultant coherence between the two SCJs would still augment credibility.

To illustrate, let us revisit the factory-farm meat consumption example. I have the initial judgment that consuming factory-farmed meat is immoral, which passes filtration and becomes a considered judgment. I then adjust my CJ set to cohere with my moral principles and background theories. My WRE set is now in reflective equilibrium. Then I investigate the WRE sets of others and find that all of our sets are highly coherent: they also have the same systematized considered judgment, and correspondent moral principles, that eating factory-farmed meat is immoral. In such a case, I can attribute my SCJs and moral principles to be more credible than before, due to this broad coherence. This augmented coherence is not the result of mere agreement, but grows from the fact that rational beings, independently following rational rules of filtration and mutual adjustment in a three-tiered system of reflective equilibrium, reached similar systematized considered judgments and moral principles. Again, we might liken this to the scientist who discovers that her filtered data set – after adjustment with her hypothesis as well as other hypotheses – coheres with the filtered data set of other scientists.

The Filtration Process versus Background Theory Coherence

Before I move on to explicate and develop the filtration process in the next chapter, it might prove clarifying to illuminate the contrast between the filtration process and background theories. The nature of the theories constituting FP and BTs is identical: both include metaphysical, normative, and scientific theories.[120] The principal difference between the FP and BTs is one of degree and not of kind: it is the degree to which the respective theories are substantiated.

If a theory is decisively substantiated, where it would be acceptable to all rational and reasonable moral subscribers, then that theory belongs as "theory filter" in the filtration process.[121] If a theory is not entirely substantiated, where it would not be acceptable to all rational and reasonable individuals who are moral subscribers, then that theory is not appropriate to filter moral judgments. By "rational," as previously mentioned, I mean individuals who are willing to submit their moral principles, considered judgments, and background theories to critical tests, and are ready to revise or discard any of the three in the face of a critical mass of counterevidence.

Considered judgments, after passing filtration, should be pitted against the relevant background theories. Though conflict between a considered judgment and substantiated background theory doesn't necessarily call for immediate discarding of a CJ, it will likely diminish the credibility of the CJ, though it still might not undermine the credibility of the CJ.[122]

---

[120] Only background theories can be moral in nature, however: the filtration process does not include moral theories as theory filters, as theory filters need to be acceptable to all moral subscribers. In one way, FP does include some kind of moral theory, or metaethical theory, in its recognition of what counts as morally irrelevant to all rational, reasonable moral subscribers.

[121] In the filtration process, the theories that filter our error-disposed judgments I am calling "theory filters." To explore the metaphor, the filtration process is akin to a complex filter constituted by several sub-filters with it, such as those complex filters utilized to make water potable in impoverished communities in the third world. These complex filters have multiple sub-filters: a textile filter, which removes clusters of bacteria; an iodine filter that eliminates several viruses, bacteria and parasites; an active carbon-filter that removes medium-sized bacteria; and so forth. In relation to MWRE, the sub-filters are the theories that bear upon whether or not an initial moral judgment can be identified as error-disposed.

[122] It's important to remember that credibility is a matter of degree, which can be represented as a numerical value between 0 and 1. There is a threshold between 0 and 1 where we might say, for practical purposes, the credibility threshold lies, above which a moral judgment is credible and below which a moral judgment is noncredible. This is similar to scientific practice where data sets might be attributed credibility ratings, even if not explicitly, noting the likelihood the data generated under error-disposed conditions. In

A considered judgment, deemed provisionally credible as it has survived the filtration process, can later lose its credibility in deference to background theories. After all, the filtration procedure is an imperfect process, as it is constituted by imperfect theory filters. While theory filters are decisively substantiated, the Popperian point forewarns that no theory is certain: many decisive and fundamental scientific and metaphysical theories have changed throughout the history of human thought. Because of this, and other factors, it's possible that a considered judgment may have been mislabeled, and may actually be an error-disposed moral judgment that passed, undetected, through the filtration procedure. Background theories function as a secondary check upon the provisional credibility of considered judgments. Imagine that a considered judgment passes filtration, yet conflicts with several background theories – background theories that are strongly corroborated just shy of the degree that they'd be included as theory filters in filtration. In such a case, it seems the additive conflict with such strong and numerous BTs decisively undermines the credibility of the considered judgment so much where the CJ should be discarded as noncredible, and thereby excluded from any WRE set.[123] In the revision and adjustment of a WRE set, if a considered judgment turns out to conflict with a number of substantiated background theories, while being supported by none, then it's highly probable this considered judgment is erroneous.

To illustrate this point, imagine Henry has the initial moral judgment "It is morally permissible to afford less moral considerability to women, as a group, than to men." This is solely based upon his belief that women are relatively cognitively inferior to men.[124] Presume there is no theory filter that decisively undermines the credibility of

---

MRE, moral judgments should never be completely discarded, but bracketed to varying degrees, proportionate to their below-threshold credibility values.

[123] This further reveals that the filtration process and background theory coherence are not different in kind, and to remind us that credibility is a matter of degree. To focus on the filtration process, if it turns out that a theory filter that filtered out certain moral judgments was a bad filter: that is, the established theory was found to be incorrect, then the moral judgments that had been discarded upon that basis should be brought back into the WRE set. Our lesson is that in MRE, we should never throw data away. Data are never 100 percent discredited or confirmed.

[124] We might imagine Henry to be an admirer of Kohlberg's (1971) initial studies on childhood development that seemed to suggest that boys reached higher levels of moral development at an earlier age than girls, where the boys would tend to become cognizant of principles and justice more readily and fully

this moral judgment. Despite the assumed lack of theory filters in FP, the basis for this moral judgment could still be challenged by several background theories: cognitive science, experimental psychology, sociology, feminist theory, and so forth. If each of these background theories, while not decisive, were highly corroborated – just short of inclusion in the filtration process – then additively, the background theories would seem to undermine the credibility of Henry's moral intuition, since the purported factual belief which serves as the basis for that moral judgment is decisively discredited by the group of highly corroborated BTs.

As exposed in this brief examination of FP versus background theories, it would be one possible option that MWRE could ditch the filtration process entirely, and just rely upon background theories to determine credible from noncredible moral judgments. Instead we could consider each background theory, determine its degree of substantiation and assign the BT a corresponding numerical value, then multiply this value, and any other BTs so represented conflictive with a moral judgment, against the numerical value of initial credibility of a moral judgment. Then a numerical value threshold for credible and noncredible could be determined, a rating above which a moral judgment would be considered credible. The theory filters, which are now being treated as decisively substantiated background theories, would have an extremely high substantiation value, which would inversely attribute a low credibility value to any moral judgment that conflicted with such theories.

In essence, this approach is just another way to look at filtration: Each moral judgment is first pitted against the most highly corroborated and relevant background theories, to determine credibility at the outset. One primary reason not to set up the methodology in this way is a practical reason: it is too burdensome and complex to keep track of every moral judgments credibility score, as well as that of the moral principles and moral background theories systematized by unchecked moral judgments of varying credibility ratings. Also, if we're taking seriously the moral judgments of other individuals, one must keep track of the generating conditions of each moral judgment, not

---

than girls. Kohlberg (1983) later revised his scoring in response to the criticisms of Carol Gilligan (1977) criticisms that his scoring method favored principled reasoning over an ethics focused on relationships and care.

only of our own, but those of others as well. By the use of the filtration process, all rational and reasonable moral subscribers can start from the same set of considered judgments, and then systematize differently – within the confines of rationality. If the filtration process were replaced by background theory coherence, there would be no initial starting set of shared, provisionally credible considered judgments, and WRE set evaluation and competition would become much more difficult and inscrutable.

One last reason to keep the filtration process as part of the method of wide reflective equilibrium, which demonstrates that it is not a mere vestige of the narrow method, is that without the filtration process all moral judgments would automatically count as considered judgments. As considered judgments, these CJs would exert pressure upon background theories and moral principles, even if they were considered judgments based upon mistaken facts or morally irrelevant bases.

Imagine, for example, that Fred believes doctors should be punished for their maliciously and wantonly inflicting pain with sharp implements upon people. Fred's intuition seems mistaken: the doctors are neither "maliciously" nor "wantonly" inflicting pain; rather they are beneficently treating patients in order to prevent future pain and suffering. Fred's intuition is based on factually mistakes and conceptual confusion.

Imagine a second case involving Fred, where Fred has the moral judgment that brown-eyed people are not morally considerable, solely due to their eye-color, whereas all other people are morally considerable. This is another crucial reason for the filtration process: it filters out moral judgments that are based upon morally irrelevant features. As explained in the first chapter, morally irrelevant features are those features that no moral subscriber would recognize as morally relevant. If no moral subscriber would validate a particular moral intuition because it is based upon a morally irrelevant feature, then this intuition should not be given entry into the arena of ethical adjudication and debate. If the filtration process were scrapped, then such erroneous moral judgments would automatically count as considered judgments, though no one would recognize them as credible.

Without the filtration process as part of MWRE, both of Fred's moral judgments – one based on mistaken facts, the other on irrelevant features – would be automatically

credible, and as such would exert pressure against other contrary intuitions that may not be so erroneous, as well as moral principles, and moral and nonmoral background theories.

To further illuminate the argument above, which asserts the need for the filtration process, consider a brief analogy to science: Fred is a chemist who generates experimental data in his chemistry lab. Imagine his starting conditions are deplorable: his beakers are dirty with chemical residue, his equipment is miscalibrated, and perhaps even some of his chemical containers are mislabeled. Fred runs his experiment anyway and generates a data set. He publishes this data set, and other scientists now must take his data seriously, at least initially, and weigh it against their own data sets (presume that their data sets were generated under no error-disposed conditions). Fred's data will likely mount a challenge to the other scientists' own data sets as well as their hypotheses.

In this way, there is a very real sense in which the filtration process needs to be the initial gatekeeper in determining what data – moral or scientific – is credible enough to be in play, and what data should be excluded at the outset from reflective revision and adjustment.

Chapter 3: "Filtration, Etiologies, and Intuition Credibility"

Introduction

In the last chapter, I examined the method of wide reflective equilibrium (MWRE) and defended it as a valid moral methodology. In this chapter, I will elucidate a crucial feature of MWRE: the filtration process.

The filtration process (FP) discredits initial judgments that arise in conditions disposed to error; these intuitions are then excluded from the set of considered judgments, which are the provisional starting points for ethical theory construction and testing.[125] Considered judgments that not only survive the filtration process, but also survive coherence tests with moral principles and background theories will be termed systematized considered judgments (SCJs).[126]

In explication and justification of the filtration process, I will appeal, via analogy, to scientific methodology. The starting point of the construction and testing of theories in both scientific methodology and MWRE is the filtration process: pruning error-disposed data from initial data to form credible data. There are several conditions that need to be satisfied, as part of both methodologies, to ensure the validity of the filtration process. In relation to MWRE, the filtration of noncredible judgments from credible judgments must be based on criteria which do not rely upon trivial conditions or presumptuous moral principles, which, as Sencerz points out, would open the filtration procedure up to charges of vicious circularity.

The primary goal of this chapter will be to identify circumstances that impugn the credibility of moral intuitions. I will suggest further expanding the role filtration serves in WRE: namely, I will examine how etiologies -- the sociological, psychological, and

[125] I am using "intuitions" and "judgments" synonymously. I employ the term "intuitions" when speaking of moral judgments generally, without distinguishing between initial judgments and considered judgments.

[126] SCJs are similar to credible scientific data which has been generated under proper experimental conditions and also cohere with corroborated hypotheses and other peripheral scientific laws.

biological shaping forces surrounding the generation of certain intuitions -- might serve to impugn the credibility of these intuitions.[127]

The Filtration Process

Filtration is a crucial process in MWRE: it separates initial judgments from considered judgments, the latter which we take to be provisional starting points of theory construction and deliberation. Since considered judgments serve to construct, support, or discredit principles, the filtration process serves a critical role in theory construction and testing.

Kai Nielson (1977) remarks that there is a dearth of discussion exploring the process of filtration, as well as a dearth of justification for proposed filtration criteria . The only philosopher Nielson cites as expounding this critical part of MRE is Tom Regan (1983). Stefan Sencerz also notes, "Surprisingly little attention has been devoted by the proponents of MWRE to this essential part of the procedure" (1977, p. 79). Part of my project will be to provide further elucidation and exploration of the somewhat neglected filtration process.

An explication and justification of the filtration process will help dispatch Brandt's objection that MWRE is a mere "reshuffling of moral prejudices" (1979, p. 22). One way of minimizing Brandt's objection is by instituting a methodology that would exclude biased intuitions from moral deliberation.[128] Generally, MWRE filters out intuitions that arise under error-disposed conditions (hereafter, EDCs). Rawls, as well as other philosophers, notes that moral intuitions may result from self-interest, self-

[127] The filtration procedure can also incorporate certain moral principles in the set of theory filters. These principles must be highly and widely corroborated: such as the principle of proportionality, which I will discuss later in the chapter.

[128] I acknowledge that there are never bias-free conditions or environments in science or human affairs – bias is always present. Nevertheless, there are degrees of bias. The condition of bias I am referring to when I mention it is where the bias is above the threshold where we believe it may prove significantly distorting. For simplicity, I only refer to "bias" as an error-disposed condition (EDC) when it is above this threshold; when it is not, I do not mention its presence, though surely some bias is always involved. Also, I acknowledge that, like the condition of emotionality, bias could possibly be an amplifying condition (e.g., a mother saving her drowning child whereas a stranger might be less likely to perform this moral act, which should be performed. Like emotionality, psychology and sociology could determine some cases where bias was distorting, when neutral, and when amplifying to credibility.

deception, historical or cultural accident, hidden class bias, and so forth.[129] I will consider several error-disposed conditions throughout the chapter.

Five suggestions for the filtration process are these: (1) ideally, initial judgments should be attributed a credibility rating.[130] (2) FP should occur repeatedly, especially when testing intuitions against principles. (3) FP should appeal to relevant and substantiated theories from the sciences, social sciences, metaphysics, and even normative theory[131] (4) FP should rely upon relatively independent criteria in order to avoid circularity. (5) FP should not rely upon trivial or arbitrary criteria in determining intuition credibility.[132]

The first point merely suggests that initial judgments, as they are subjected to FP, should be attributed varying degrees of credibility, rather than a bimodal attribution.[133] The second suggestion emphasizes that FP should not be applied once, but repeatedly in order to double-check intuition credibility, especially in cases where there's something at stake, such as the testing of moral principles or ethical theories. The third point demands the filtration process incorporate as "theory filters" scientific, social scientific, and substantiated metaphysical theories in order to determine credibility determination; one example I will examine later in this chapter regards three recent psychological studies of the basic emotion of disgust and its influence on moral evaluation of actions: namely, that the emotionality of disgust can dispose moral intuitions toward error. The fourth point

---

[129] I will examine only those EDC candidates which I believe to be the most promising and illuminating. Other EDC candidates, as Rawls mentions, such as "hesitation," I find to be terminally problematic, and will not examine. Please note that the mention of an EDC candidate does not equate with endorsement of it as an EDC.

[130] Brandt suggests attributing to each initial judgment a credibility rating ranging from 0 to 1. Initial judgments with a low credibility rating are filtered out from the set of considered judgments, whereas initial judgments with a high credibility rating are included in the set of considered judgments (1979, p. 20).

[131] Such theories might include psychology, moral anthropology, sociology, cognitive psychology, evolutionary psychology, theories of human action, theory of human identity, and so forth. Normative theories might include the theory of proportionality, principles of harm, the role of morality in society, etc.

[132] One way filtration criteria might be trivial is if particular criteria is deliberately selected in order to "cook the books," where we select criteria particularly so they help us reach the principles we want to reach.

[133] Though for practical purposes, we may treat intuitions as either credible and thus accepted into the set of CJs, or noncredible, and rejected from the set. Science makes a bimodal distinction as well, in terms of practically accepting some data sets and rejecting other sets, in theory formation and testing, though these data sets would deemed as having varying non-bimodal degrees of credibility.

will engage the objection of circularity, and show that FP can avoid circularity if it employs nonmoral filtration criteria that are independently substantiated, or if it relies upon normative principles which are highly and widely corroborated[134] and are sufficiently discrete from the intuitions which they are assessing. The fifth and final point will show that the filtration process can and must avoid relying upon trivial criteria in intuition credibility determination.

Before launching into explication of the last three points,[135] it may prove helpful to briefly revisit the moral methodology of MWRE and its parallel to scientific methodology.

Filtration and Scientific Practice

The filtration procedure (FP) parallels the controlled conditions present in scientific experiment and theory construction. In scientific experimentation, the conditions and parameters of the experiment must be strictly set, scientific equipment must be precisely calibrated, and interference must be minimized. In fact, even the scientists' mental demeanor or professional disposition might possibly come into consideration regarding error-disposition.[136] If an experiment is discovered to have been performed under conditions disposed to error, such as from lack of instrument calibration, the data set is discarded: this data set is not used to derive, corroborate or test

[134] Highly corroborated refers to the number of experiments. Widely corroborated refers to the different "angles" from which this corroboration occurs: different experimental setups, different cultures, etc. This is relatively parallel to the example of an ornithologist corroborating "All ravens are black" by examining one million ravens in America (which would be a high corroboration) versus an ethnologist corroborating "all ravens are black" by examining a dozen ravens on each continent of the globe.

[135] I believe the first two points are relatively uncontroversial, so I will omit them from discussion.

[136] A scientist's perception of the world can be "theory-laden." For instance, a geocentric astronomer would view certain planets as having irregular orbits that occasionally exhibited a backward "retrograde" motion, whereas a heliocentric astronomer would view these planets as being occasionally outstripped by the Earth's orbit around the Sun. Likewise, if a pre-Darwinian scientist were to view variations of finches in nature, he would view them as imperfect deviations from the perfect mean, whereas a neo-Darwinian scientist would view them as specialized animals that were progressively improving in fitness relative to a niche. Similar to theory-laden concerns, another problem in science is "confirmation bias." In the history of science, there are numerous examples of scientists who operated under a confirmation bias, attentive to corroborating evidence and blind to counterevidence. There are even scientists who practiced purposeful bias, such as scientists who purposefully "cooked" their data, as they self-interestedly sought prestige.

a hypothesis. In psychology, for instance, experimenters frequently conduct controlled double-blind studies to avoid the generation of biased data.[137] In chemistry, chemists ensure that certain experimental conditions are kept constant – such as volume, pressure, temperature, and so forth – if these conditions could interfere with experimental results. In addition to external conditions, there are internal cognitive conditions of the scientists themselves. Margaret Holmgrem explains:

> "In science, theories that have been developed provide grounds for believing, under certain circumstances, either that an individual lacks the capacity to make an accurate observation or that he is liable to make inaccurate observations until he has been made aware of the disorienting condition that affects him. For example, these theories explain why a terribly nearsighted person does not have the capacity to make certain visual observations, and why a colorblind person is liable to be mistaken in his observations of color, at least until he has been made aware of his defect" (1987, p. 111).

In a similar way to scientific practice, the filtration process of MWRE is in place in order to exclude those initial judgments which were made under conditions of error. As previously stated, Rawls cites some possible conditions of error:

> "We can discard those judgments made with hesitation, or in which we have little confidence. Similarly, those given when we are upset or frightened, or when we stand to gain or lose one way or the other can be left aside. All these judgments are likely to be erroneous or to be influenced by an excessive attention to our own interests" (1979, p. 47).

Norman Daniels supplements this list by indicting judgments resulting from self-deception, historical and cultural accident, hidden class bias, and inadequate information (1979a, pp. 265-268). Peter Singer adds intuitions that are vestiges of discarded religious systems, or of social and economic customs that were necessary for survival in the past (1981, pp. 490-527). It isn't immediately clear if every item proffered by these philosophers as an error-disposed condition is justified. In the course of this chapter and the next, I will examine a few of these conditions and attempt to provide some justification for their inclusion in the filtration process.

The aforementioned EDC candidates suggest that initial moral judgments could be error-disposed if they arise under such conditions. If a moral judgment does arise under error-disposed conditions, this does not necessarily mean that the moral judgment

[137] This typically is the case in pharmaceutical studies and drug trials. Oftentimes, a double-blind study isn't possible; nonetheless, it is an ideal condition that is instituted when possible.

is incorrect; it merely implies that the moral judgment lacks credibility, and, subsequently, we should not rely upon such judgment as a provisional starting point of moral methodology. In our science analog, a set of data could in fact be correct (and we might even see that the data set conforms with other reliably-generated data sets); nonetheless, if the data set in question was generated under error-disposed conditions, that data set must be discarded.

EDC specification includes but is not limited to the following general constraints: they must be precisely formulated[138], have defined scope[139], and be context sensitive.[140] These constraints can be discovered and substantiated through appeal to empirical theories.

We commonly acknowledge error-disposed conditions as we morally navigate our ordinary lives. If we have a moral judgment, and suspect that some error-disposed conditions are present, we often refrain from moral deliberation or seek to "massage" our intuitions.[141] One common error-disposed condition we recognize is emotionality.[142] In some cases, emotionality seems to unduly influence our moral judgments. Empirical research can help determine the occasion and degree of the influence various emotions have upon our judgment. I discuss one empirical study below, which investigates the emotion of disgust and its putative effect upon moral evaluation.

[138] "Emotionality," for instance, should be rarified into subcategories: for instance, acute resentment should be discrete from ebullient happiness.

[139] The parameters of error-disposition must be specified: for example, emotionality might tend to distort one set of moral intuitions (e.g., regarding punishment) but not another set (e.g., regarding duties to self). Note: these parenthetical examples are purely speculative.

[140] An EDC can be error-disposed under one set of conditions, but not under another set: for example, retributive intuitions in a hunter-gatherer society, where cooperation is essential to survival, may be less error-disposed than in a society where goods are more bountiful. This criterion wouldn't necessarily court moral relativism. Virtues, for instance, are partially context (or culturally) sensitive, without being entirely relative.

[141] One way to "massage" one's intuitions is to consider the adage, "What if I were in another person's shoes?" This is supposed to help one overcome bias or spur sympathy by providing a different perspective than might be myopic in its perception.

[142] Emotionality can also be a credibility-amplifying condition, which I will discuss later.

Disgust

In illustration of how the empirical sciences can help us identify error-disposed conditions, consider recent psychological studies regarding the basic emotion of disgust.[143] When human subjects are disgusted, this emotionality appears to heavily influence their moral assessments.

One study, conducted by Thalia Wheatley and Jonathan Haidt (2005), presented all subjects with post-hypnotic suggestions to feel a pang of disgust upon reading the arbitrary word "take" or "often." When subjects read a variety of hypothetical moral scenarios, those subjects presented with scenarios employing the term "take" or "often," (as opposed to some alternative word, not hypnotically suggested for) would judge the actions of the characters in the scenario as more morally wrong than would the control group. The control group's scenarios were identical, except for the presence of these words.[144] This means that these arbitrary words, which triggered disgust in the subject, heavily determined the degree of the subjects' negative moral evaluation. This suggests that disgust represents an error-disposed condition, at least in certain circumstances.

One objection to this conclusion, of course, is the contention that these words, though trivial, actually contributed toward verisimilitude, where the disgust-amplified moral condemnation of the characters in any given scenario were of an appropriate degree. This is similar to the case of viewing the factory-farming video, where the disgust upon viewing animal maltreatment might actually liberate a person from their entrenched species-biases.

There are two responses to such an objection. First, we must acknowledge a distinction between morally relevant features and trivial features. For instance, in the factory-farming video case, our disgust is elicited by the pain and suffering of sentient animals; morally relevant features are eliciting our disgust reaction, which in turn affects

[143] Disgust is considered a basic emotion according to "Disgust," (Rozin, P., et al., 2000).

[144] The scenarios involved characters performing acts where no overt harm was occurring: (1) breaking a promise to your grandmother to visit her grave after she's dead (2) Using an American/Brazilian flag to clean the bathroom (3) A family eats its pet dog after it's hit and killed by a car (4) a brother and sister romantically kiss on the lips (5) A man masturbates using a dead chicken, and then cooks and eats it.

our moral evaluation of the practice involving those animals. If the moral question at hand concerns if meat eating is morally permissible, presumably knowledge, not only propositional but experiential, of animal suffering is relevant to the ethical determination. After all, the ethical question is *about* animals and their treatment. In the hypnosis case, however, the words "take" and "often" are not *about* the content presented in the moral scenarios. Therefore, "take" and "often" cannot be morally relevant features; they are trivial features of how the scenarios are presented.

Consider a hypothetical hybrid case, where the post-hypnotically-suggested subject didn't watch a factory farming video depicting animal suffering, but instead read the passage, "John *often* eats meat" – as opposed "John *frequently* eats meat." If the subject had been given the post-hypnotic suggestion to feel pangs of disgust when confronting the word "often," the subject might generate the correct moral intuition that John's practice of often eating meat as more morally wrong, as opposed to the more moderate moral assessment of the control group who viewed the sentence that didn't have the trigger. Presuming for the sake of argument that eating factory-farmed meat is significantly immoral, John's disgust-amplified moral intuition happens to achieve a greater verisimilitude than the control group; nonetheless, his verisimilitudinous intuition is less credible that the intuitions of the control group. This greater verisimilitude is achieved due to the efficacious yet arbitrary feature of the word "often" present in his scenario.

In parallel, briefly entertain a thought-experiment, where a subject is hypnotized with the post-hypnotic suggestion that every person wearing a brooch is a morally corrupt person. It might just so turn out that this correspondence obtains in the real world: all morally corrupt persons happen to wear brooches. Irrespective of this coincidental correspondence, the fact remains that the presence of a brooch is not a feature that has any relevance to the state of the person's moral character.

A second psychology study extends the investigation of the first. In this second experiment, subjects were presented *non-moral* scenarios involving the disgust-trigger-words "often" or "take." The scenario describes a student council representative who

"often picks" or "tries to take up" topics of broad interest of discussion.[145] Many subjects evaluated the characters' action in the scenario as "somewhat wrong," while two subjects assessed his behavior as highly wrong. Asked for explanations of their moral evaluation, subjects remarked, "It just seems like he's up to something," or "It just seems so weird and disgusting" or "I don't know [why it's wrong], it just is." If disgust is triggered in response to the presence of morally irrelevant features, such as the arbitrary presence of the words "take" or "often", and the elicited disgust strongly influences the subjects' moral intuitions, even in cases where there is no moral violation occurring whatsoever, then we can question the credibility of the resulting intuitions. This experiment suggests that disgust, when it is based on morally irrelevant features, disposes us to error in our moral intuitions. Furthermore, the experiment shows our post-hoc rationalization of the erroneous moral intuitions, which have been artificially triggered via the basic emotion of disgust.

Disgust, however, needn't be an error-disposing condition in cases where that disgust is elicited by morally relevant features: for instance, when the disgust is elicited by animal suffering.[146] Disgust can be amplifying in the case of the factory-farming video, when the disgust arises from sympathetic recognition of suffering, which represents a morally relevant feature.

Disgust can be an error-disposed condition not just in artificial cases, as in the case of post-hypnotic suggestion, but in "natural" cases, where disgust arises because it has been evolutionarily or sociologically selected for, but arises from morally irrelevant features.[147] For instance, disgust might arise in circumstances where there were good evolutionary reasons to have disgust reactions: for example, when the disgust reaction motivated the avoidance of objects that were harmful. A person might have negative moral evaluations of diseased individuals, we might speculate, thereby exiling them from

[145] The summary of this experiment comes from Greene's 2007 article, "The Secret Joke of Kant's Soul" (p. 58).

[146] The example of disgust parallels a case presented in chapter 4, regarding moral evaluations regarding trust, when trust is elicited by the arbitrary injection of oxytocin in subjects, rather than when oxytocin is secreted in response to "normal" visual and olfactory cues.

[147] For illustration purposes, I provide only a cursory discussion of two examples of "naturally occurring" disgust, here; I will present a more in-depth investigation and analysis in the next chapter.

a community not just due to their ill-health and possible contagion, but because they were seen as possessing a polluted moral character.

Another example of disgust aversion naturally arising is the nearly-universal aversion to incest.[148] Disgust toward sexual relations between brothers and sisters has a strong influence upon moral intuitions condemning this practice. In evolutionary history, this disgust was predicated upon morally relevant features: namely, birth deformities, lower genetic resistance to disease, and lessened reproductive fitness. In contemporary industrialized societies, where birth-control is available as well as medical preventions and treatments for disease, the evolved disgust aversion to incest no longer has the same underlying morally relevant features as a possible justification for moral condemnation of the practice.[149] In fact, when subjects are presented with an imaginary scenario involving incest between adult brother and sister, and are then asked why they are opposed to the incestuous sexual relations occurring in the case, from which no harm results, the subjects cannot provide reasons for their moral condemnation; they are, as Haidt describes them, "morally dumbfounded" where they often laugh at their peculiar inability to provide reasons.[150]

Consider another empirical experiment involving disgust, which doesn't involve post-hypnotic suggestions but rather involves environmental stimulus to trigger a disgust reaction in subjects (Schnall, Haidt, Clore, 2004). In this study, subjects were told to sit at a desk and fill out a questionnaire regarding a variety of moral scenarios. Those subjects seated at a disgusting desk – a desk that was stained, sticky, and located near a trashcan overflowing with grimy pizza boxes and dirty-looking tissues – responded with significantly more negative moral evaluations to the presented moral scenarios than the

[148] I examine the incest example more thoroughly in chapter four. For now, the discussion is merely cursory.

[149] There may be other auxiliary reasons to be against incest, but this wouldn't sufficiently explain the strong disgust aversion to incest deeply and pervasively present across cultures and human history.

[150] Another example where disgust seems to generate moral intuitions of condemnation is homosexuality. James Rachels examines the usual suspects for homosexuality condemnation and finds them to be insufficient and usually *ad hoc* (1986, ch. 1). There seem to be no morally relevant features upon which to predicate this moral condemnation. Disgust, whether socially or biologically influenced, seems to influence our moral judgments. I would speculate that there is some significant correlation between condemning homosexuality and feeling that homosexual relationships are "disgusting." This would be an interesting subject of psychological study and experiment.

control group, not situated at disgusting desks. As in the hypnosis case, the artificially elicited disgust had significant influence on the negative moral evaluations of the subjects in the scenarios.[151] The disgusting environment was external and unrelated, however, to the moral dimensions presented in the scenarios, which subjects were morally evaluating. In this way, the disgusting environment of the desk is a morally irrelevant feature to the moral evaluation of the content presented in the moral scenarios.

Disgust, as a basic emotion, appears to have a significant degree of influence upon moral intuitions. There seem to be legitimate reasons to accept some disgust-driven moral intuitions while rejecting others, contingent on the moral relevancy of features to which the disgust is respondent. It would be of further interest to investigate moral intuitions in other cultures, as well as our own, that seem closely tied to disgust. Haidt (1993) investigates the basic emotion of disgust in relation to cultures around the world, and identifies four moral bases that seem present in human cultures. One of these bases he terms "purity and pollution", which is a moral base deeply entrenched in some cultures whereas absent or not prominent in others. Haidt seems to characterize all four bases as equally valid. I am inclined to disagree with Haidt, as I suspect that an investigation of this moral base might show that sometimes a culturally inculcated disgust aversion to some objects was based upon morally irrelevant features. Such evidence could undermine claims of validity, and show the moral judgments based on such occurrences of disgust were noncredible.[152]

Relevant Cognitive Conditions

Error-disposed conditions contrast with those conditions necessary for moral judgments to be deemed credible, which Sencerz calls "relevant cognitive conditions" (hereafter, RCCs). One RCC would be the condition of being informed of the relevant facts surrounding the focus of the intuition. For instance, if Abe sticks Ben with a needle, Ben might have the intuition that it would be retributively appropriate to retaliate

[151] The subjects in this experiment, it should be noted, were those that identified themselves as highly sensitive to their own bodily states.

[152] This topic is one I'm interested in pursuing, but is beyond the scope of the current project.

against Abe; however, if Abe is a doctor administering a beneficial inoculation, Ben's initial judgment to retaliate would be inappropriate.

Further relevant conditions during intuition generation include conceptual clarity, rationality, and impartiality.[153] The agent generating intuitions must understand situational context, possess a capacity for rationality, and not be biased via self-interest.

Consider the formal condition of conceptual clarity. To be a credible intuition, the individual having the intuition must be clear on the relevant concepts. Regan (1983) presents the example of euthanasia. If an individual has an intuition that euthanasia is wrong, we cannot assess whether or not that intuition is credible until it's clear that the individual understands what euthanasia is, and what is involved in the controversy. The same would be true of capital punishment, abortion, among several other contemporary moral issues. Cultural relativism would assert that "conceptual clarity" is lacking by ethnocentric individuals who hastily condemn certain practices of other cultures. The cultural relativist (not necessarily the ethical relativist) would defend some cultural practices as morally correct, and assert that objectors misunderstand the cultural concepts. For example, scarring the face of young boys in certain tribes in Africa[154] might cause a cultural outsider to morally evaluate this cultural practice as immoral. However, if that outsider came to understand that this cultural practice was not an instance of wanton cruelty to children, but rather a revered rite of passage to manhood, which was eagerly accepted by the boys, and was a sign of beauty and respect, then that individual (having this newfound conceptual clarity) would likely slough off his previous condemnatory intuitions, and could now intuit this cultural practice as morally permissible.

In addition to factual accuracy and conceptual clarity, a third formal condition is the criterion of rationality. Regan (1983, p. 128) narrowly fills out this condition as one of consistency: "To fall short of the ideal moral judgment by committing oneself to a contradiction is to fall short as one possibly can." Regan illustrates this criterion in reference to abortion. If a man judges that all abortions are morally wrong, but then

[153] Sencerz calls these conditions "formally correct" conditions (1986, p. 84).

[154] Ritualized "scarification" is fairly common in parts of West Africa and New Guinea. It represents heightened social status and spiritual "completeness."

judges that there was nothing immoral about his wife's abortion, then these two judgments, all else being equal, are rationally inconsistent. The man must seek to resolve this inconsistency.[155]

This criterion of rationality could be filled out further. One addendum might be a formal principle of equality: that rationality – as well as justice – demands a similar judgment about how similar individuals should be treated. This is closely related to the principle of consistency, above. In addition, rationality might include certain psychological conditions, such as an accurate perception of one's self and one's surroundings (so we can exclude Descartes' mad men who believe they are made of glass, or are luxuriously clothed when actually unclothed); this is another way of stating the formal condition of factual accuracy (or conceptual clarity). Also, rationality should include a willingness to subject one's own views to critical scrutiny, and revise or reject principles and intuitions when they fail to stand up to this scrutiny. This is similar to rationality in science, where the scientist, to be rational, is willing to subject her hypothesis to critical testing – trying to falsify the hypothesis – and being willing to reject the hypothesis and previous data if it fails to stand up to such severe tests.

Manifestations of rationality might be specified further beyond the principle of consistency, which I will subsequently count as the first of four rationality principles. I would suggest the following three principles to be addendums to the rationality criterion[156]: (2) the principle of fairness: treating like things alike (3) the principle of proportionality: retributive action should not exceed the degree of the offense (4) the principle of maximization: all other things being equal, it's rational to maximize positive values and minimize negative values: for example, it's rational to choose the action that

[155] We might expect the man, faced with this inconsistency, to realize that he's being biased toward his wife. He might, alternatively, end up changing his stance on abortion, given his experiential exposure to moral salient features that often surround decisions to terminate pregnancy. He might come to understand these features, given his intimate relationship with his wife. It might be a practical difficulty determining which of the two contradicting intuitions need to be "massaged," but it is clear some revision is necessary. An individual in this situation might be able to resolve this inconsistency by investigating and appealing to some error-disposed condition under which one of the two intuitions arose.

[156] These four rationality conditions are normative, which might invite possible charges of circularity; I respond to this possible charge directly in a later section. These four rationality principles are internal to the filtration process, and should be kept distinct from the notion of rationality, external to the filtration process, that refers to theory construction, testing, and revision. The latter sense of rationality is related to Duhem's thesis (1954), and is discussed in chapter 2.

minimizes rights violations, or maximizes positive consequences, depending the moral outlook to which the person is subscribing. I will further examine rationality-based principles in subsequent sections.

Impartiality is a fourth relevant cognitive condition that Sencerz mentions, comprising the set of relevant cognitive conditions. Contrarily, bias is an error-disposed condition. Bias should be contrasted with partiality. In cases of friendship, for instance, the partiality we should give to a friend over strangers is arguably justified. I take the impartiality standard to be more specific and constrained than the prescription that we "treat everyone alike." The impartiality condition relates back to the rationality criterion. For instance, impartiality requires that you not arbitrarily grant yourself or associates a higher intrinsic status, or provide exceptions that you wouldn't allow for all. This means, for instance, that you are only warranted to give favor to a friend over a stranger as long as you would consistently allow others to favor their friends over strangers to a similar extent and in similar contexts.[157] The essential undergirding concept is rational consistency, which will be examined further shortly.

Primarily, violating impartiality constitutes bias when it violates institutional or social rules, roles or responsibilities in which one participates. Again, impartiality requires that we not make ourselves or associates an exception: for instance, a policeman should not let his friend off the hook for a crime, when he would not condone other officers acting in the same way toward their friends.

Also, as previously stated, we shouldn't grant our friends or ourselves more *intrinsic* morally considerability than other individuals: one counts for one and only one. Nevertheless, given our special relationship to friends, we can attribute them more *extrinsic* moral considerability, as long as it is in accordance with rational consistency.

Bias, whether toward ourselves or our family or friends, is not always readily apparent, however. Self-interest is one persistent form of bias, and is often frequently subconscious.[158] We do acknowledge in our personal lives that under certain

[157] It's possible that, despite the fact you would countenance consistently extending the allowance of friend-favoring to others as well, that such a practice in some cases might be immoral (e.g., nepotism in hiring). The point here is to provide a negative constraint, not a sufficient positive test.

[158] Psychology might be able to provide a basis to determine when self-interest is present, and when this

circumstances, our self-interest may bias our intuitions. For instance, when asked to judge a certain situation where our personal interest is at stake, we may recuse ourselves from normative deliberation, citing our bias (e.g., "But don't listen to me; I'm probably biased."). For example, as a white male, I might question the credibility of my initial normative intuitions regarding certain socioeconomic issues (e.g., affirmative action in college admission) that bear upon diminishing the advantages I have. Because of my self-interest, I might have the disposition to be hostile toward or hastily dismissive of claims that I enjoy "white male privilege" and thereby should be amenable to some redistribution toward the underprivileged. I should at least be wary of the credibility of my intuitions in such circumstances, and take measures to examine the situation and circumstances more closely and carefully, and try to familiarize myself with different points of view, rather than faithfully relying upon my bare initial intuitions. To consider another example, we can understand a member of a hiring committee recusing himself from participating if it turned out that one of the applicants was a good friend of his. Even though the committee member could strive to be impartial – and not be aware of any bias on his part, or think such bias surmountable – it seems reasonable for him to recuse himself – just in case – as he realizes any bias might unfairly give advantage to his friend at the unjustified disadvantage of the other job applicants.[159] Psychology and sociology are resources that can be appealed to in determining conditions of bias. For example, sociological studies can identify bias, not just on an individual scale, but a more general societal scale: one example is how physical appearance is often a biasing factor in hiring, employment and promotion.[160]

---

self-interest may be distorting to moral intuitions. There are several recent psychological studies suggest human beings are biased and unaware of such bias. One example is doctors' bias towards pharmaceutical companies who provide perks, such as paid vacations, elegant dinners, and even trivial amenities, such as delivered fast-food lunches or free pens and so forth.

[159] On the other hand, he might be in the special position to bring to light certain relevant facts concerning his friend, the applicant, such as the applicant's reliability and good character. He might provide such relevant facts and his personal evaluation of his friend to the committee members, for their consideration. It isn't these additional facts that constitute the bias, however; it is rather one's judgment which is compromised (though frequently this compromised judgment determines what is accepted as the facts in the first place).

[160] For instance, studies in sociology suggest that taller men get promoted faster; attractive women are more likely hired in positions where appearance is arbitrary; etc.

One additional RCC that I would add to the cadre of relevant cognitive conditions is MRF sensitivity. For an individual's intuition to be credible, the individual must be cognitively sensitive to morally relevant features, as opposed to trivial features. For instance, bias might make an individual morally insensitive to certain morally relevant features, as previously discussed. For example, a doctor, biased by a pharmaceutical company's gifts, might be more apt to overlook morally relevant features in a patient's profile in order to prescribe that patient a sub-optimal drug from the favored pharmaceutical company.

Rationality, Normative Laws, and the Circularity Objection

While rationality principles, delineated above, must recognize moral concepts, such as morally relevant features, this acknowledgment will not beg any of the questions at hand. The proposed rationality principles included in FP are not culprits in circularity because they refer only to normative assumptions which are independent of the moral principles and considered judgments under review. Furthermore, these normative assumptions might be considered "normative laws" by the fact that they're highly and broadly corroborated to such an extent as they appear acceptable to all moral subscriptions.[161]

To clarify, consider again an analogy between MRE and scientific methodology: considered judgments parallel observational/experimental data, moral principles parallel scientific hypotheses, and normative laws parallel scientific laws.[162] In science, a hypothesis is tested against the background assumption of numerous scientific laws, which are, as a body, presumed to be true.[163]

[161] I am referring to these normative assumptions as "normative" instead of moral because the assumptions reflect what, I argue, are inexorable norms general and necessary to all moral subscriptions. I refer to the normative assumptions as "laws" in an analogy to science, as laws are highly and widely corroborated, and thereby constitute the background conditions in testing hypotheses, and are typically not the subjects of contestation themselves.

[162] The differences between the trinity of data, hypotheses, and laws are not one of kind, but of degree: namely progressively increasing abstraction and corroboration.

[163] At least presumed true by the scientist, if not the philosopher.

In illustration, consider the famous Millikan oil-drop experiment (1909), which proposed to measure the electric charge of the electron. This experiment presumed the correctness of gravitational laws, as well as laws concerning electric fields. Given these assumptions, the charge on the oil droplet could be determined. These assumptions, however, had already been independently established, and as such were background assumptions upon the shoulders of which Millikan could test his hypothesis.[164] It may turn out that certain laws are incorrect, which would invalidate the hypotheses which relied upon these laws as presumed background assumptions; this is quite a different charge from vicious circularity, however.

In a similar way to scientific methodology, the rationality criterion, instantiated as four rationality principles delineated above, relies upon certain theoretical assumptions; in this case, the rationality criterion relies upon the assumption that some features are morally relevant whereas others are not. Nonetheless, this reliance upon moral assumptions is independent of the intuitions under consideration. That is, while the moral component of the rationality criterion may be theory-laden in that it recognizes moral concepts, the principles that include these concepts will not be specifiable to one particular theory, but to morality in a wider sense that all moral subscribers would accept.[165] For example, to assess whether the principle of fairness is satisfied, this may require the acknowledgement of certain features, such as sentience or well-being, as morally relevant. This acknowledgement of a feature as morally relevant is not dependent upon any particular moral principle or any considered judgments; the question concerning moral relevance is separate from the question at hand.

In review, relevant cognitive conditions are those conditions necessary for intuitions to be deemed credible. If intuitions do not arise under relevant cognitive conditions but, rather, arise under error-disposed conditions, those intuitions will

[164] That hypotheses are always tested in bundles, rather than as isolated singularities, relates to "Duhem's Thesis," a problem in philosophy of science, which points out that if an experiment does not bear out the hypothesis' prediction, the scientist need not scrap the hypothesis, but can scrap any part of the bundle of hypotheses serving as background assumptions. This isn't always rational to do, however: for example, if to save his favored hypothesis, a scientist scrapped the law of gravity, or some other highly corroborated theory. This problem is explored in more depth in chapter 2.

[165] "Moral subscribers" denotes those individuals who accept that moral propositions are consistently and objectively true or false, and who subscribe to an intelligible ethical form of life.

contrarily lack credibility. Below, I provide a summary list of error-disposed conditions (EDCs) and provide a brief example of each condition in turn.

Error-Disposed Conditions

A condition is error-disposed if any of the following is the case:

(1) The condition influences the subject to be blind or insensitive to certain features (MRFs) that are morally relevant to the intuition.
- For instance, anger is an EDC if it tends to make one insensitive to the fact that a person who harms one may have done so unintentionally.

(2) It causes the subject to generate a moral intuition that is not predicated upon any morally relevant features (but is predicated upon morally irrelevant features)
- An agent's disgust toward an object might result in her negatively evaluating the object in moral terms, though the object is morally neutral
- A human being's bias toward his own interests might motivate him to rationalize that since animals are furry, they mustn't be deserving of moral treatment.

(3) It causes the agent to alter her moral judgment due to morally irrelevant features.
- An increase in an agent's disgust toward an object might result in her increasing her negative evaluation of the object, even though the cause of the additional disgust is unrelated to the object of moral evaluation[166]

(4) It interferes with one's understanding of factual accuracy.
- For instance, if a police officer is fearful, full of adrenaline, he may be prone to mistakenly interpret a suspect to be holding a gun rather than a wallet, and thus mistake the suspect as a morally appropriate object of preemptory violence.

(5) It interferes with conceptual clarity.
- Reactionary indignation might cause a pro-life advocate to ignore the conceptual distinction between *genetic* human being and *moral* human being.

(6) It violates principles of rationality.
- The subject morally judges the act of cutting in line to be immoral when it's done by other people, but believes it's morally permissible when he cuts in line himself

[166] Take for example the post-hypnotic suggestion hybrid case where "often" or "take" increased the subject's disgust and, subsequently, increased their negative moral evaluation of it that. While this increase might be veridical to moral truth, this increase is not justified as it arises from morally irrelevant features.

(7) It leads to any other EDC obtaining
    - Intense adoration, as an instance of emotionality, may result in the adorer to factually evaluate, unreliably, that the adored is superior in all faculties and abilities to those of all others, which leads to the adorer providing special favor to the adored

(8) It causes an agent to have an unjustified degree of positive or negative evaluative moral judgment toward an object
    - A woman who feels contempt at being cut in front of in line may have the moral judgment that just retribution permits hitting the offender with her car.[167]

In terms of any possible charges of circularity, conditions (1) through (7) do not seem particularly problematic. Conditions (1) through (3) were considered in previous examples, such as in regard to disgust as an EDC. Conditions (4) through (6) seem to be objectively determinable through empirical means. Condition (7) seems a trivial condition, expressing that some conditions should be considered error-disposed not because they directly affect moral judgments, but because they lead to other error-disposed conditions which do directly dispose moral judgments toward error. The most controversial condition of the set seems to be condition (8), which appears to tempt charges of circularity. After all, upon what basis can we judge someone's moral intuition to be deficient or excessive in degree, without first relying upon our own moral intuitions as a corrective? I take up this question in the section below, with specific attention to retributive moral judgments.

A House is not a Gnome: Excessive Retribution

Consider the following example:

> A neighbor deliberately breaks one of Beth's garden gnomes. Beth is very angry. In what she feels is justified retaliation, Beth burns down her neighbor's house. When

[167] This would be in violation of the principle of proportionality, as previously described as part of the rationality criterion; this is described further, below. Though this example may seem initially outlandish, it is based on a crime report from 2007.

> interviewed, Beth concedes that sure *now* it seems to her that her initial moral judgment was excessive. But she maintains that, back then, her moral judgment assessed that burning down her neighbor's house was a just retributive desert. She shrugs, and queries upon what basis we may deem her initial anger-influenced moral judgment as excessive or miscalibrated. Perhaps, she continues, the correct moral judgment was her initial "excessive" response, and her calmer moral judgment, which she has now, is in fact deficient and incorrect.[168]

How are we to answer Beth's challenge? That is, upon what non-circular basis can we say that Beth's action was excessive and unjust rather than appropriate and justified? We cannot merely cite that her initial intuition doesn't comport with the intuitions the rest of us have. And we cannot merely state, sans independent justification, that moral intuitions generated during heightened states of anger are less morally reliable than moral intuitions generated during moderate states of calm.

Before delving into a direct answer to Beth's challenge, it should be noted that the circularity charge doesn't arise if any of the seven criteria for EDCs, listed above, indict Beth's intuition. It seems somewhat difficult to imagine a person with an excessive moral judgment that passes all seven EDC criteria delineated above. But presuming that Beth's intuition did not arise under the error-disposed conditions (1) through (7), I will argue that we can still assert that her moral judgment was excessive, without being guilty of vicious circularity.

First, it is important to understand what *would* count as vicious circularity. One example of vicious circularity would be if we deemed Beth's anger-influenced intuition as noncredible simply because other neighbors had moral judgments different from hers: namely, judgments that prescribed more moderate vengeance, against which Beth's judgment seemed excessive. And while multiple agreements may, at times, count as better evidence for a position being true, it certainly is not always the case, and mere agreement is not the same as substantiation.

---

[168] I'm presuming that both moral judgments for the same situation cannot be correct, given our presumptions of objective morality, as previously defined. Moral propositions are objectively and consistently either true or false. This includes propositions regarding degrees of rightness and wrongness. Concerning the above example, her second calmer moral judgment is either more or less verisimilitudinous than the former moral judgment in regard to its approach to objective and consistent moral reality.

This attempt at justification counts as circular because what is trying to be established is that an intuition exceeds an appropriate threshold, yet this verdict can only be delivered by first presuming that a threshold has been already substantively established. For instance, a bystander of Beth's arson might reason, "Beth's intuition is objectively excessive because is exceeds my own intuition." But Beth's intuition cannot be deemed as excessive without first assuming that the bystander's intuition is not excessive. You cannot determine what is excessive in the first place without relying upon some kind of independent substantiation. In brief analogy, imagine a layperson is in a room with several paintings – some are fakes and some are authentic period-pieces. The layperson is given the task of determining authenticity by merely examining the paintings. The layperson cannot deem one painting a fraud, just by showing it is different from a majority of the other paintings. The layperson might indict one anomalous painting, declaring "This is a fake!" She might be then asked how she came to this determination, to which she might respond, "Because it is different from the others." But then again how does she know the others are authentic? In both cases, it is trying to be established that a certain instance departs from the standard by already presuming that some standard is already substantively established.

How, then, can we deem Beth's moral judgment as excessive without inviting charges of vicious circularity? One way we can do this is by appealing to the concept of proportionality. The principle of proportionality prescribes that in exacting retribution, one is limited to, at most, an eye for an eye – rather than two eyes for one eye.[169] The principle of proportionality is a moral principle; nevertheless, its inclusion as a theory filter in the filtration process does not lead to vicious circularity; I will discuss this later in the section.

Before I delve into the principle of proportionality more abstractly, it might help to provide one initial argument. Imagine if scientific studies, similar to those investigating disgust, were to suggest that anger has a tendency to cause agents to make retributive evaluations in amoral cases where there were no initial trespass or offense:

[169] All other factors being equal. Additional factors that justify exceeding an eye for an eye might be the harm caused by the initial action, such as shattering one's sense of serenity, taking advantage of one's unreadiness, or considerations of deterrence of future attacks.

that is, in cases where there were no morally relevant features present. If retributive intuitions arose in amoral cases, due to the presence of anger as a determining causal factor, then it would seem we would have some reason to extrapolate, in similar contexts where the same degree of anger is present: that is, we would have reason to suspect that retributive moral judgments may be excessive in cases when the trespass or offense, while not entirely absent, is slight. Expressed in the metaphor of *lex talionis*: If anger causes a person to morally judge that she may take an eye when no eye has been taken – which is certainly disproportionate – we have reason to extrapolate that she may be similarly mistaken when she morally judges that she may take two eyes when just one eye has been taken. This extrapolation seems reasonable, independent of a presumptive reliance upon the principle of proportionality. Empirical evidence suggests that emotionality does in fact significantly influence our moral judgments. As we have already examined via recent psychological studies involving the emotion of disgust, disgust seemed to dispose the moral intuitions of experimental subjects toward error in both amoral and moral cases. Unfortunately, I'm not currently aware of any moral psychology experiments focusing upon the emotion of anger and its influence upon our moral judgments. Nonetheless, the point remains that empirical research could in fact substantiate this point in a way that is independent and thereby non-circular.

The principle of proportionality, more abstractly considered, would also determine Beth's retributive judgment to be excessive. While the principle is itself a normative principle, it is not merely a collection of considered judgments. It is a principle that is highly coherent with other WRE sets: disjunctive sets of considered judgments, moral principles, and background theories found in other individuals and cultures. Proportionality is a pervasive moral principle found throughout moral traditions spanning history and cultures.[170]

[170] Normative theory filters in FP are supposed to be broadly acceptable and uncontroversial (this is one way theory filters differ from background theories). One objection to the inclusion of the principle of proportionality as a normative theory filter in FP is that act utilitarianism might reject it, as act utilitarianism does not recognize retributive desert. If all else is equal – namely, presuming that utility in a case will not be maximized by rewarding/punishing disproportionately – the principle of proportionality does not seem one to which act utilitarianism would have reason to object, even while they may not actively endorse it. For this reason, the principle of proportionality seems acceptable, even under the act

While a highly and widely substantiated principle, the principle of proportionality is still a *moral* principle, of sorts, which might invite charges of circularity: namely, utilizing a moral principle, which has in large part been constructed by moral intuitions, to evaluate the credibility of other moral intuitions. However, the employment of the proportionality principle is not viciously circular, as the principle – and its accordant systematized considered judgments (SCJs) – not only are highly and widely corroborated themselves, but also cohere with a mass of other moral principles and moral judgments, furthering overall coherence at several levels. Due to this total high and wide corroboration, the principle of proportionality might be considered a normative law rather than a moral principle.[171]

In clarification that any circularity is not vicious, consider the Millikan experiment once more in comparison. The intent of the experiment was to measure the charge of an electron, and yet the charge could only be determined by first presuming certain scientific assumptions as settled: gravitational and electric forces, specifically. One background assumption of the experiment was the conservation of energy: that the energy in a closed system would stay constant. Absent this background assumption, the measurement of the charge of the electron would not be determinable.

In a way, the question of electric charge is not entirely discrete from the issue of the conservation of energy: both concern electrical charges and forces. Nevertheless, the law of the conservation of energy is so highly and widely corroborated, that it is considered a scientific law, which has been independently established irrespective of any assumptions or dependency about the charge of the electron.

Relating this science example to the case of Beth, above, we can compare our employment of the principle of proportionality in criticizing her initial moral judgment as excessive in the same way we might employ the conservation of energy to criticize a data point as excessive. In the Millikan experiment, for instance, we might utilize the conservation of energy to determine if there was a flaw in the experimental setup: for

---

utilitarian view, as a normative principle to be included in FP. A second point is that even if any theory filter is discovered to be unacceptable, MWRE allows for revision in light of this change.

[171] Again, this parallels the difference between experimental hypotheses and scientific laws, which are a difference in degree rather than kind: nonetheless, the difference is still significant between the two.

instance, if there appears to be more energy in the closed system after the reaction rather than before, we would discard the data point as noncredible. In a similar way, we might utilize the principle of proportionality to determine if there was a flaw in Beth's relevant cognitive conditions: for instance, if there appears to be disproportionality between the original offense and her moral judgment regarding proper retribution, we would discard Beth's moral judgment as noncredible.

Much like the normative principle of proportionality, the law of conservation of energy is just a highly corroborated hypothesis itself, as it relies upon other previous data sets generated in experiments like the current one, and yet it is also *widely* corroborated by generated data under other conditions, in other sciences, and coheres with other scientific laws.

In addition to the principle of proportionality, a second method to determine whether if Beth's intuition is excessive is to test her retributive judgment in deference to the principle of consistency, which is another principle under the rationality criterion. We can ask Beth, presuming she's reliable, how she would morally evaluative if the positions were switched. Would her generated intuition endorse the same degree of retribution, this time against her own interests, to be appropriate or excessive? Testing an intuition in this way is, of course, limited by one's imagination, experience, and is still vulnerable for bias. Nonetheless, there still might be ways to validly determine whether a person is adhering to the rational criterion of consistency.[172]

Credibility-Amplifying Conditions

In the same way that eight error-disposed conditions, listed previously, can be distilled from our discussion, eight credibility-amplifying conditions likewise can be distilled. Consider the following eight conditions to be the reverse side of the credibility coin, delineating credibility-amplifying, as opposed to error-disposed, conditions:

[172] One noted problem with this, however, is how to determine – once we arrive at this consistency – which moral judgment is credible. Presumably, given individuals' prevalent bias toward self-interest and rationality to justifying the pursuit of self-interest at the expense of others, we might suspect the initial retributive judgment, such as Beth's, to be the less credible judgment. This likely would necessitate further argumentation, however.

(1) The condition causes one to predicate their intuition upon MRFs.
- Experiencing disgust when observing, first-hand, the squalid and dehumanizing conditions of slave workers, the subject generates the moral intuition that slavery is immoral.

(2) It makes one cease or diminish basing their intuition upon morally irrelevant features.
- Greater conceptual clarity when observing the facial-scarring tradition causes one to cease his moral condemnation of the practice predicated upon the disfigurement of the boys' faces.

(3) It causes the agent to alter her moral judgment due to morally relevant features.
- An increase in an agent's disgust toward factory farming, due to experiential observation of it first-hand, might result in her increasing her negative evaluation of factory farming. In this case, the cause of the additional disgust is related to the object of moral evaluation.

(4) It enhances factual accuracy.
- Partiality of a mother to her son allows her to sympathetically understand that the motivation behind her son's act of vandalism was a feeling of helplessness rather than malice; subsequently, she assesses him as less blameworthy for his act than an outsider.

(5) It enhances conceptual clarity.
- Emotional calmness allows one to consider all sides of the euthanasia controversy, and realize that both positions exhibit a respect for life.

(6) It enhances adherence to rational principles
- Reflection upon the inconsistency between his condemnation of cutting in line by others and his justification when he, himself, cuts in line leads the agent to revise his initial judgments in deference to this rational inconsistency.

(7) It prevents or inhibits another EDC from manifesting.
- A subject is in the habit of considering "how would I see it from their perspective" in moral deliberations, which has a tendency to diminish personal bias by considering other points of view.

(8) It causes an agent to have a justified degree of positive or negative evaluative moral judgment toward an object
- An agent might reward a neighbor, whom he doesn't particularly like, with a Christmas fruitcake, when he reflects that the neighbor contributed significant effort and skill into neighborhood upkeep and improvement.

Emotionality as a Credibility-Amplifying Condition

One widely recognized source of error-disposed intuitions concerns immoderate emotion. Regan (1983, p. 129), for example, asserts that "the hotter (the more emotionally charged) we are, the more likely we are to reach a mistaken conclusion, while the cooler (the calmer) we are, the greater the chances that we will avoid making mistakes."[173] This caveat seems reasonable, *prima facie*, as we can easily imagine situations where emotionality seems to distort a moral intuition.

Emotion needn't play a distorting role in the generation of intuitions, however. In fact, emotion might lend to *greater* clarity of moral intuition generation. Daniels (1979a, p. 258 n3) remarks: "Sometimes anger or (moral) indignation may lead to morally better actions and judgments than 'calm'.". For instance, inflamed emotion may help overcome calcified prejudice and engender intuitions of appropriate empathy.[174] For example, viewing a graphic video depicting the maltreatment of animals might result in emotional disgust and sympathy, which overruns whatever embedded species-bias one might have, and results in generating intuitions condemnatory of factory farming: namely, that we should not exploit animals for trivial luxuries, but acknowledge that they deserve humane treatment; whereas, before one's intuitions might have judged that animals were "obviously" not worthy of much moral consideration.

I speculate that error-disposed conditions most often obtain in cases where emotionality overruns rationality, or rationality overrides emotionality, though I won't provide any argument here.[175] We grant initial credibility to intuitions because they arise from what we presume to be credible faculties: namely, our rational and emotional

---

[173] His parenthetical statements.

[174] By "appropriate empathy," I am not trying to beg any questions, but am speaking conversationally: I am relying on our common notions of what is appropriate and inappropriate regarding sympathy. However, if I were to analytically delve into analyzing a notion of "appropriate empathy," I would suggest that human beings have not only rational capacities but sympathetic capacities as well, and that these latter capacities can be at least roughly bounded. For example, being emotionally unmoved by the suffering of a loved one suggests a deficiency in the same way as not following or acknowledging the rational rule of consistency suggests a deficiency. Jonathan Bennett (1974) illustrates this idea in his examination of Himmler and Huck Finn, which I discuss later in this section.

[175] I am not suggesting here that emotionality is irrational, quite the opposite: Both faculties are essential parts of our overall moral faculty. I imagine that this current discussion could be appropriately related to Hume's sentiment-based ethical outlook, which might be an interesting connection to explore. I won't explore this possibility in this project, however.

faculties. These moral faculties need to be kept in balance. One way credibility is undermined, I conjecture, is when one or both of these faculties is deficient or excessive, or when one of these moral faculties is subjugated under the dominance of the other.

Jonathan Bennett (1974) explores this topic through character studies in fiction as well as history. In regard to history, he examines the Nazi, Heinrich Himmler, and suggests that Himmler was grossly deficient in his emotionality – particularly sympathy/empathy – though he was not necessarily deficient in his rational faculties. Bennett's suggestion is that Himmler's rational faculties and emotional faculties were not in balance, and that the latter was enslaved to the former.[176]

In contrast to Himmler, Bennett examines the fictional character of Huckleberry Finn (1885). Bennett details how Finn's strong feelings of close friendship with the runaway slave, Jim, help Huck overcome his principled morality and reach the correct ethical decision to continue to help Jim rather than turn him in. It appears that Huck not only fails to realize he's doing the morally correct action, but believes he's behaving immorally. I would argue, however, that Huck is, in fact, motivated by a sense of ethics, in the face of the conventional and religious morality of the time. After reading a letter from his teacher, Miss Watson, he considers that his decision may determine his place in either heaven or hell. Tearing up the letter, Huck declares: "All right, then, I'll go to hell" (1885, ch. 31, para. 25). Huck is taking a stand of personal moral conviction against the morals of his society.

In Bennett's article, Huck Finn is characterized as a fictional hero. Huck serves as a moral exemplar insofar as Huck's moral faculties seem to be more properly balanced than they would have been had he unquestionably accepted the "conscience" cultivated by his society. Huck has been told by his society that Blacks are inferior and not fully human. Huck's sympathy and own practical rationality is able to overcome these culturally-inculcated prejudices, and he is able to discard society's instilled moral

[176] In the elucidative contrast, we might consider an example of excessive sympathies paired with deficient rationality. For example, imagine a faction of animal liberation activists that trespass and vandalize a cancer research laboratory in order to liberate a few laboratory mice: we might consider their sympathies to be excessive and their rational faculties to be deficient in such a case. We can imagine their actions would injure the animal liberation movement; be harmful to the animals they 'liberated'; and impede valuable cancer research.

judgments for new moral judgments in sensitivity to the morally relevant features he observes during his developing association and friendship with Jim: for instance, in observance of Jim having full human faculties: intelligence, sentience, dignity, hopes, emotions, and so forth.

This example illustrates that emotions can have an illuminative effect, in addition to the distorting effect Rawls charges, where intuitions that arise under certain emotional conditions are *more* credible than if they would be if they arose in absence of such conditions. For example, if a person feels sorrowful, it might make them more likely to be sympathetic to the plight of animal suffering mentioned in the previous example. Such illustrations of emotionality as sometimes constituting a credibility-amplifying condition, rather than as an error-disposed condition, is a valuable point. It conveys that an account of emotionality as an EDC demands precise specifications. For example, the emotional conditions of being "upset or frightened" need to be distilled. There may be a specifiable difference, regarding error-disposition, between being "emotional" in an empathetic way versus being "emotional" in an enraged way; the two might diverge in regard to credibility, where only the latter is an EDC, whereas the former could be a neutral – or even amplifying – condition under which intuitions arise.[177] Nonetheless, the fact that emotionality can assist, rather than inhibit, legitimate considered judgment generation does not establish any critical difficulty with FP. Rather, such objections merely helpfully highlight the practical difficulties in the development of the filtration account, thus far. These difficulties can be overcome, by and by, with the mortar of empirical research: most specifically, further studies bearing upon moral psychology.

Error-disposed conditions in scientific methodology also demand specification; there are multitudes of notable accounts that seek to do this.[178] In regard to moral methodology, we are left with the question as to how to determine those conditions under which intuition-generation is error-disposed. These conditions need to be explored and

[177] I'm ill-equipped to delve into psychological analyses of the effects of different emotions on the fidelity of intuitions – factual as well normative intuitions – but there appears to be an increase in of such related studies.

[178] E.g., Hacking, 1983; Mayo, 1996.

specified. Social scientific disciplines, such as psychology, sociology, cognitive neuroscience, moral anthropology, among other fields, might help us identify EDCs.

Without delving into empirical research, it seems that in our everyday lives, we often acknowledge that our emotional conditions may dispose us toward bad moral judgment. Indeed, we recognize that part of maturity and rationality is an awareness that our emotional states affect our moral attitudes and behavior, and require us to be deferentially cognizant of this fact when relying upon moral judgments that arise when we are certain emotional states that may dispose us to error.[179]

Etiologies and Error

Moral judgments tend to be significantly influenced by sociological, psychological, and biological factors which may affect the credibility of that intuition.[180] In order to determine the credibility of a moral intuition, the etiology of the intuition should be considered. An etiology is the causal story or causal conditions determining a moral judgment.[181] Descartes (1641) provides a paradigmatic example of etiology as an error-disposed condition. Employing a hyperbolic method of doubt, Descartes presents, but ultimately rejects, the possibility that God could have made him in such a way where is his prone to error, "But if it goes against his goodness *to have so created me* that I am always deceived, it seems no less foreign to it to allow me to be deceived sometimes…" (p. 76, para. 5). He then goes on to reject that it is God who may be deceiving him, positing rather that it may be some evil demon that is responsible for leading him epistemically astray. Under the supposition of an evil deceiver, Descartes concludes he cannot be sure of even mathematical truths, such as that two and three equal five, or that the sides of a square do not exceed four.

[179] For example, a parent might be careful when doling out punishment to her children after enduring a particularly aggravating day at work.

[180] An etiology of an intuition might impugn an intuition's credibility on the basis of its predisposition to violate a RCC: for example, if an individual was victim of some severe childhood trauma, we might expect that some subset of their intuitions might be disposed toward irrationality. Psychology might provide specification of this subset.

[181] Richard Joyce (2006) deems the causal origins behind an intuition a "genealogy," concerning where the intuition came from.

I do not wish to throw MWRE to the wolves of Cartesian skepticism. I do wish to show, however, that Descartes' evil demon argument is illuminative in conveying that etiology matters in credibility determinations. And although the etiology Descartes invokes impugns the credibility of *knowledge* claims – not moral claims – this same template seems applicable to moral intuitions as well. Working under the assumption of moral objectivity, moral judgments are to serve as provisional epistemic claims. Intuitions are provisional starting points in the apprehension of moral knowledge. For such a reason, I would argue that the two – knowledge of facts as well as knowledge of provisional moral facts – are analogous.

A thought experiment, similar to Descartes, might illustrate how etiologies can constitute a class of EDCs. Imagine you undergo psychological experimentation.[182] When you wake up, you are told you've been subject to intensive hypnotic suggestion as well as neurological brain-tinkering by two evil, yet honest, scientists: a hypnotist and a neurologist.[183] The scientists inform you that they worked together to implant all sorts of haphazard moral intuitions in your mind, just to see what you'll do with them.[184] The hypnotist did this through powerful subconscious suggestion whereas the neurologist did this by tweaking with your cognitive physiology and brain chemistry. Given the experiment, it seems reasonable to question the credibility of your subsequent moral intuitions – certainly in comparison to your set of intuitions before the experiment. One reason we might question credibility is because the deceiving scientists haphazardly implanted intuitions into you, thereby entirely bypassing all of those faculties which we take to be truth-tracking.[185] In this way, the *etiology* of your intuition set can diminish the credibility of that set.

---

[182] I'm presenting the thought experiment in 2nd person perspective in parallel to the way Descartes' presents it.

[183] The hypnotist is to represent sociological forces of suggestion/indoctrination; whereas, the neurologist is supposed to represent biological forces, such as evolutionary bias, which I will discuss in the next chapter.

[184] This thought experiment is based upon the thought experiment originally presented by Robin Hanson (2002) in her paper on health care ethics. Coincidentally, this thought experiment is quite similar to the one presented by Richard Joyce (2006, ch. 6), which I discovered upon the publication of his book. Joyce has us imagine we have swallowed a "belief pill" that causes you to believe "Napoleon lost Waterloo."

[185] Again, assuming moral objectivity. MWRE assumes that our moral faculties, or whatever faculties our intuitions arise from, are capable of being truth-tracking. Similarly, we ordinarily believe that our sense

We might imagine the meddling scientists informed you that they limited their experiment to implanting intuitions in regard to particular moral categories. For instance, they might inform you that they only meddled with intuitions in regard to retribution; this would leave you to doubt the credibility of intuitions in those categories, while not other normative categories.

Social Etiologies and Credibility

The above thought experiment suggests how certain etiologies can undermine the credibility of related intuitions. Not all etiologies undermine credibility, however. Consider, for example, intuitions that arise from a thoughtful and comprehensive moral education, an education that incorporates looking at different perspectives, investigating morally salient reasons for actions, weighing consequences while being mindful of relationships and duties. This etiology might *amplify* rather than discredit the credibility of an initial intuition. For example, if an individual has the initial moral intuition that slavery is immoral, we might trace this intuition back to her moral education. Her moral education is the reason she has this intuition; it cultivated empathy in her for others, and it taught her extensive facts about the exploitive, degrading nature of slavery. In this way, this etiology wouldn't seem to undermine this particular intuition's credibility; in fact, an investigation of the etiology seems to amplify the intuition's initial credibility. An initial intuition could be said to be amplified when, upon investigation of its generation, its etiology includes relevant cognitive conditions: such as relevant facts, conceptual clarity, rational capacities, and so forth. An intuition's etiology could satisfy these conditions in varying degrees. An intuition might be said to be amplified when its etiology is one that rises above the threshold of acceptable credibility.[186]

---

faculties, where our perceptions are generated from, are truth-tracking. Likewise, the scientific methodology to which most practicing scientists subscribe (at least the scientific realists) presumes that experimental equipment is truth-tracking, and isn't just producing random or trivial phenomena without verisimilitude to reality.

[186] Acceptable credibility is where the degree of credibility necessary for an intuition to pass the filtration procedure and be accepted as a considered judgment. This closely parallels the degree of credibility a data set (generated by a scientific experiment) needs to have in order for it to be included in construction or testing of a hypothesis. We might say a scientific data set is "amplified" when it not only

In examination of etiologies, let us consider the following example: A politician who is a recognized war veteran has the moral intuition that war is immoral – or at best a necessary evil. It's likely he has this intuition due to his past experience. The etiology, or causal story, determines or at least heavily influences this intuition; however, it doesn't undermine the credibility of this intuition. On the contrary, we might consider the veteran's intuition far more substantiated because of its etiology; in fact, we might consider it more substantiated than the very *same* intuition generated by some other person lacking that etiology – for example, a politician who lacks any firsthand experience with war.[187]

In contrast to the war veteran, we can consider the case presented by Sturgeon (1992), who considers the possible moral intuitions of a young man in relation to war. The young man has a coherent NRE set: his considered judgments about war match his principles regarding war. Yet his initial intuitions seem to arise under conditions that are not formally correct: that is, the young man does not have accurate or sufficient factual information regarding what war is. According to Sturgeon's example, the man has only seen pro-war propaganda, whereas he is not cognizant of certain realities of war. Knowing that this young man's etiology involved propaganda, we would likely think this etiology would diminish the credibility of his intuition. Propaganda involves overemphasis of some facts, deemphasis of counter-facts, and distortion of concepts. In this way, propaganda violates a formal condition criterion in the filtration process: namely, by the subject not possessing a correct apprehension of the facts or concepts to generate a credible intuition.

---

satisfies the threshold of credibility, but also surpasses that threshold: that is, that the data set was under conditions which were tightly controlled and there is high confidence that there were no factors that could interfere with the fidelity of the results.

[187] These etiologies need further specification, however, as we might imagine the war general to inappropriately glorify war, have been indoctrinated with false reasons as to why his country engages in war: e.g., the ideology of spreading democracy rather than economic interests. This example merely serves as ways in which an etiology could be relevant to intuition credibility.

It should be noted that an error-disposed etiology only discredits an individual's intuition; it does not show that that intuition is false.[188] In fact, there might be other etiologies supporting that intuition.

Overdetermination

Merely citing one questionable source of a particular intuition is not sufficient to undermine the credibility of that intuition: an intuition could be causally overdetermined. To briefly delineate, an etiology diminishes the credibility of an intuition if the following conditions are the case:

(a) the etiology suggests the intuition arose under error-disposed conditions
(b) there are no other known etiologies that significantly determined the intuition (even partially, which would result in partial credibility), which would legitimately support the intuition[189]

Criterion (b) might seem difficult to satisfy: how can it be exhaustively established that there is no other corroborative causally determining factor? This seems a practical impossibility. Nonetheless, the burden of proof doesn't demand we exhaustively establish the absence of some other determining etiology. If, for instance, we know an individual's moral judgment is sufficiently determined by his own self-interest, and after diligent search we conclude that no other morally relevant causal etiologies are evident, then we can reasonably conclude his intuition is noncredible.[190]

In regard to overdetermination, consider an example of a woman who grew up being reared by parents who were animal activists. It would be unfairly hasty to dismiss

[188] Freud's psychological etiology for the belief in God as an emotional coping mechanism, for instance, would only, at the very best, serve to discredit some basis of belief in God; it wouldn't serve as any argument *against* God's existence.

[189] It might be impossible to exhaustively establish that no other etiologies are present which would validate an intuition as credible. Nevertheless, after a diligent search, if we find that no candidates seem viable, the burden of proof is at least shifted upon the critic to provide such a validating etiology.

[190] There might be corroborative nonmoral explanations that could support the intuition that slavery was *extrinsically* immoral: for instance, that the practice negatively affects the slave-owner in the same way that executing human beings might have negative effects upon executioners.

her animal rights views as mere indoctrination. Her intuitions could *at the same time* be a product of impartial and comprehensive reflection. To state in another way, her intuitions are overdetermined if she would have been an animal rights advocate, anyway, by mere parental indoctrination, and, if at the same time, it turns out that she has rigorously investigated different points of view, examined the relevant factors, and has discovered morally relevant reasons independent of any indoctrination. If the latter is part of the etiology of her intuitions concerning animal rights, then her intuitions would have arisen under relevant cognitive conditions, and would thereby be credible. In this case, her intuitions couldn't be undermined merely by citing the etiology of parental indoctrination, as her intuition would remain credible due to the alternative, legitimate foundations supporting it.

In this way, in assessing the credibility of an intuition, it's necessary not only to investigate if there's an error-disposed etiology, but also if there are any other legitimate etiologies that can be recognized as sufficiently determining the intuition.

However, it's not sufficient simply to show that there is support available for a given intuition to be grounded upon; the individual's intuition has to be *actually* grounded upon that support. In illustration of this point, Sturgeon (1992) presents the example of the "militant abolitionist" who wants to abolish slavery. The abolitionist has the intuition that "slavery is intrinsically immoral," yet the etiology of this intuition is one of self-interest, as he is resentful that others are prospering in the slavery trade, and he is left behind (whereas, in fact, he'd likely have the intuition that slavery was morally permissible, were he benefiting from the institution). Having determined this etiology, which indicates the intuition arises from self-interest, we can conclude that the intuition is not credible. Certainly, an intuition that slavery is intrinsically wrong *could be* based on several compelling reasons; and surely the intuition that slavery is intrinsically wrong is in fact correct. Nevertheless, the etiology of the militant abolitionist's intuition is error-disposed; and the moral conviction of the abolitionist is actually rooted in self-interest and self-deception.

Examining the militant abolitionist example further, if we discovered that his intuition was based on a kind of class-envy, and the intuition really had nothing to do

with the intrinsic features to slavery (such as exploitation, suffering, inequality, oppression, and so forth), then the intuition is noncredible since it's based upon morally irrelevant features. The reason the militant abolitionist has the moral intuition that slavery is intrinsically immoral is a reason completely unrelated to the morally relevant features of slavery: namely, that he is not benefiting from the practice while others are. His intuitions, then, do not arise from the relevant facts. We could imagine that if we spoke with the militant abolitionist about the reasons he felt slavery was wrong, the reasons he gave in condemnation of it might seem odd, in that they were extrinsically focused not on the features of slavery itself, but upon its place in the socio-economic fabric (for instance, how it results in unfairness and classism). We might imagine the militant abolitionist to be so self-deceived, however, that when he talked to us about his intuition that slavery was wrong, he would talk about the cruelty, subjugation, exploitation, and just finish by saying something like "it just seems wrong to me." In this case, it would ostensibly seem that his intuitions arise from the right features about slavery: namely, intrinsic features.

We might deem this slightly different character the rationalizing abolitionist. The features he cites, even if they are morally relevant, are just rationalizations; they are, in fact, causally inefficacious to his condemnation of slavery. We can consider a counterfactual to illustrate the true etiology of his intuition: namely, if the "rationalizing abolitionist" were economically benefiting from slavery, he would no longer believe slavery to be wrong. This counterfactual shows that the intrinsic features of slavery are not the conditions that are generating the abolitionist's intuitions: in fact, we might as well replace "slavery" with another practice and we'd likely expect the rationalizing abolitionist's position to remain steadfast in condemning the practice, X, from which he was not benefiting. Of course, in outward justification, he'd have to focus on features of practice X, and he might in fact happen upon features that seemed to be morally relevant features; nevertheless, his intuitions would not be credible due to the fact his vocalized reasons were not causally related to the generation of his intuition condemning the practice of slavery.

Consider a revised version of our militant abolitionist, "the enlightened abolitionist." The enlightened abolitionist has the same intuition as the militant

abolitionist that slavery is intrinsically immoral. Originally, this intuition sufficiently arose from self-interest, as well as self-deception (regarding the "intrinsic" nature of the intuition). After witnessing slavery firsthand, the militant abolitionist becomes aware of certain features of slavery which support his original conclusion: slavery is intrinsically immoral. The intuition "slavery is intrinsically immoral" is only credible after the abolitionist becomes aware of features that relate to the intrinsic features of slavery; until that point, his intuition lacks credibility. As soon as such supporting reasons are introduced, namely the fact the intuition is predicated on morally relevant features, the original intuition gains credibility from that support.[191] In such a case, the intuition generation is "overdetermined:" That is, the intuition is determined by more than one source, and each source is sufficient for its presence.[192] It doesn't matter that, in our case, it was the self-interest of the militant abolitionist that caused him to investigate the features of slavery in the first place – subsequently actually finding supporting reasons, which he internalized to become the enlightened abolitionist. How he got there does not matter concerning credibility; all that matters is that the reasons are indeed supporting.[193]

However, these reasons need to be *sufficiently* supportive in the following way: one determining reason that generates the intuition that slavery is immoral is due to observation of the intrinsic features of slavery; that is, these intrinsic features should be the supporting cause of this intuition about its intrinsic immorality. The credibility of the intuition is proportionate to the degree that the intuition arises from supporting causes: in this particular case, the intrinsic features of slavery.

At the same time, there can also be additional determining factors present: in this specific case, the abolitionist's self-interest could be *sufficient* in determining the

[191] Note, morally relevant features needn't generate any particular intuition: that is, Abe can heed the intrinsic features of slavery and generate the intuition that slavery is morally good, rather than morally wrong. As long as his intuition is arising from these morally relevant features – the intrinsic features of slavery that we find as rational beings acknowledge *not* to be irrelevant (such as self-interest) – then the intuition is credible so far. The intuition may still be subject to examination in relation to *other* etiologies that might affect credibility, such as factual error: for example, if the subject looks at the intrinsic features of slavery and generates the intuition that it is morally good because Blacks deserve it because they're sinful descendants of Cain who murdered his brother, Abel. I believe there would be several resources that would show this person's beliefs to be factually incorrect.

[192] Though they needn't be supportive of the intuition's credibility; in this case, one is supportive while the other is not; it could be the case that neither is supportive

[193] By "supporting" here I mean that the features his intuitions arise from are morally relevant features.

intuition – just as long as the self-interest is not *necessary* for the effect. In the case of the enlightened abolitionist, self-interest is causally sufficient for the effect; nonetheless, cognizance of the intrinsic features of slavery is also sufficient for the effect. As long as there is *some* legitimate (that is to say, not error-disposed) foundations that the intuitions are based upon, those intuitions are supported, and thereby credible. Some inquiry could be made into what original etiology precipitated the intuition generation in the first place (e.g., self-interest); however, this original etiology does not discredit those intuitions as long as at some point another legitimate set of supporting reasons is introduced. Perhaps the second etiology "supplanted" the first etiology, in terms of justification. We might imagine both etiologies as columns perpendicular to the roof of the intuition, where only the second etiology is supportive (and the first is a false, or shaky, column, or no column at all). Even though the fact the abolitionist *originally* generated the intuition that slavery is intrinsically immoral because of self-deceptive self-interest does not mean that the intuition can be later supported through a different etiology (e.g., examining the suffering and exploitation of slavery firsthand). As we have seen, noncredible intuitions can be substantiated by later etiologies.[194]

Again, we can understand this case of the enlightened abolitionist in the form of counterfactual: If slavery was in fact in the self-interest of the enlightened abolitionist, he would still generate the intuition that slavery was immoral. If it would be the case the he would *not* generate this intuition, if slavery were in his self-interest, then his current intuition that slavery is immoral would not be credible.

To clarify the conditions for intuition credibility, it might help to consider a counterintuitive case. Consider the case of the scientific slavery advocate, where this individual bases his intuition that slavery is morally permissible based upon the scientific evidence available at the time. This slavery advocate has examined the scientific evidence that has been presented by ostensibly credible sources. Alleged "intelligence tests" performed on Blacks and Whites, indicate that Whites are significantly more

[194] The Steven Spielberg movie "Schindler's List" (1994) portrayed this sort of supplanting of reasons. Schindler started hiring Jews to work for him based on his self-interest (profit), but progressively this self-interested motivation is supplanted with a moral motivation. Both etiologies might be sufficient to motivate him to do the morally correct action, but only the latter etiology is sufficient to justify his actions as "moral."

intelligent than Blacks. Evolutionary theory of the time suggests that Whites are higher on the evolutionary chain than Blacks, and so forth. Given this scientific evidence, a slavery advocate generates the intuition that slavery is morally permissible because Blacks are significantly inferior to Whites.[195] In this case, we would have to say that the slavery advocate's intuition that slavery is morally permissible is credible. The reason it is credible is that the intuition arises under relevant cognitive conditions: for instance, it is in deference to the best scientific facts of the time. Of course, the science of that day may have not been credible, as it appears to have been significantly influenced by bias and prejudice. Certainly, scientific evidence today indicates facts to the contrary. Nevertheless, given this individual's understanding of the purported facts which frame his interpretation of the features of slavery (e.g., that it is not human exploitation or subjugation, but more on par with the prudential harnessing of farm animals for labor), we will be forced to accept that this intuition, during that time, was credible, even though false.[196]

We might imagine other cases in which an intuition arises from multiple etiologies, two (or more) of which lend *partial* credibility. For instance, I might have the intuition that wantonly kicking dogs is immoral. Imagine my etiology for this is twofold: (1) I'm have some evidence to believe dogs may have feelings, though I'm not sure (2) Even if they didn't have feelings, I have some evidence to believe the wanton abuse of animals fosters vices in human beings.[197] Both etiologies lend some partial credibility to this intuition; together, they would seem to lend more credibility than they would singularly.

---

[195] For the example, we can assume that the individual has had limited exposure to Black persons, and has encountered no counterfactual evidence to the contrary of what the science asserted.

[196] Again, we might impugn the credibility of the scientific process that resulted in evidence that supported racial inequality. If the science providing the facts is noncredible, the resulting moral intuitions based upon those noncredible findings is also noncredible (presuming some one was in a position to determine the scientific evidence was noncredible). However, if the scientific findings were made under the proper scientific conditions, and unbiased scientists just happened to have gotten it wrong, then the moral judgments would have been credible at that time. Given contemporary scientific evidence to include into FP, we can now rerun those racist moral judgments through FP and find them noncredible, as they run afoul of factual accuracy as a relevant cognitive condition.

[197] Kant argues on this basis in his Lectures on Ethics, stating that cruelty to animals may lead to cruelty towards men, though he maintains that animals lack inherent moral considerability (2000, p. 240).

In brief review: it is not sufficient for credibility that the intuition *happens* to be correct; credibility depends on the conditions of intuition generation, and whether those conditions were error-disposed. An agent could generate intuitions from highly error-disposed conditions that coincidentally also happen to be correct. This intuition, however, is not credible. Secondly, it is not sufficient for credibility that the intuition *could* be solidly grounded on other reasons not in evidence. For instance, if the militant abolitionist has the intuition that slavery is intrinsically immoral, and this intuition entirely arises from self-deceptive self-interest, then it's irrelevant that the intuition *could* be solidly grounded upon strong supporting reasons. In order for an intuition to be credible, it needs to arise *from* those reasons. Credibility of intuitions concerns the actual causal etiology, not just a possible etiology that isn't in fact responsible for the generation of that intuition.

Etiologies and Amplification

Sturgeon (1992) states that nonmoral explanations can also *amplify* the original credibility of an intuition. Consider the example of an individual who generates the intuition that eating meat is immoral. This intuition might have been formed as a result of legitimate philosophical reasoning by the individual. This intuition is deemed credible as it wasn't formed under error-disposed conditions but was formed under relevant cognitive conditions. This intuition might gain some further credibility, however, if it surpasses the minimal criteria for qualifying as a considered judgment. For instance, if a moralistic vegetarian leaves his synthetic-leather armchair and visits factory farms and witnesses the cruel treatment of animals firsthand, with this direct experience, the vegetarian's originally credible intuition would be further supported by greater satisfaction of RCCs: the intuition is further informed by facts via firsthand experience, and his conceptual clarity of factory-farms is augmented.[198]

[198] A distinction could be made here between two kinds knowledge: referential knowledge and experiential knowledge. Knowing experientially provides additional facts: namely, "what it's like." For example, the abolitionist might at least have a better idea of what slavery was like. This might engage his faculty of sympathy much more than any armchair reflections on the subject of slavery.

In a similar way, we might understand that a non-self-interested abolitionist -- who has never directly witnessed the suffering and exploitive nature involved in slavery – might still have a credible intuition that slavery is wrong; nevertheless, the abolitionist who has directly seen slavery in action would have an *amplified* credibility to their intuition (in comparison to the previous abolitionist). We might re-imagine the abolitionist as having only direct experience with slavery, but having not philosophically reflected upon the practice in any critical and analytical way. His intuition might be already credible, as it didn't arise under error-disposed conditions, yet we can understand that the initial credibility would be *amplified* if it additionally arose after the aforementioned reflection. Indeed, we'd find the intuition initially credible from direct experience *or* philosophical reflection, but we'd find the two factors together to lend the *most* credibility to the intuition that slavery is intrinsically wrong. This is because the intuition is predicated upon a deep understanding of the facts: namely, increased propositional as well as experiential knowledge of slavery – the "what it's like" knowledge of slavery. In addition, the intuition becomes predicated upon a deeper understanding of the concepts involved in slavery: that it undermines autonomy, causes suffering, decreases well-being. In this way, nonmoral explanations can amplify credibility instead of undermining credibility.[199]

In another example, Sturgeon presents the case of a man who has the intuition that homosexuality is immoral (1992, pp. 97-98). Sturgeon explains that the nonmoral explanation/etiology of this intuition is that the man has unacknowledged fears about being homosexual himself. Presuming this nonmoral explanation is true, the intuition arises solely from fear rather than any facts about homosexuality. In this way, the nonmoral explanation seems to undermine the credibility of the intuition.[200]

---

[199] Emotional generation isn't sufficient for credibility, however. Consider the sympathy we might have for insects being killed, or trees "bleeding" sap from the savage blow of a man's axe. There need to be *some* facts to vindicate whether or not the emotional response is appropriate or not. Contrariwise, facts alone often don't lend the charge to normative intuitions as some direct or analog experience do. Additionally, without some kind of sympathy and experience of feeling it doesn't seem likely we could have normatively evaluative intuitions at all.

[200] It's not in virtue of the explanation being nonmoral that it is diminishing to credibility; it's the kind of nonmoral explanation that is involved in generating the intuition. This should become clear through the examples following.

Consider a second case, where a man has the intuition that pedophilia is immoral. We might imagine, similar to the case above, that there is a nonmoral explanation of this intuition – namely that the man has unacknowledged fears about being a pedophile himself. Presuming that this nonmoral explanation is true, the intuition might arise from fear rather than facts about pedophilia. If this were the case, that nonmoral explanation (if exhausting the reasons why he thinks pedophilia immoral) would, too, seem to undermine the credibility of the intuition. We would imagine that there would be other reasons a man would have the intuition that pedophilia is wrong: for instance, the intuition that it's wrong to inflict pain and suffering on innocent children. If these reasons were the source of the intuition, as well as the source of this man's fear that he was a pedophile, then this intuition would have credibility to the extent that the intuition arises from these reasons. The aforementioned facts about pedophilia bear upon its moral value; fear of pedophilia by itself, on the other hand, does not seem to bear upon moral value.

Chapter Conclusion

The filtration process is a critical procedure of the method of wide reflective equilibrium. FP determines the credibility of our initial moral judgments, which results in the acceptance of considered judgments as provisional moral data, from which we construct and test moral principles. This chapter fills out the internal machinery of the filtration process, through relevant cognitive conditions, error-disposed conditions, and credibility-amplifying conditions. This internal machinery needs to be grounded in such a way that it avoids circularity and arbitrariness. The filtration process can avoid circularity if the filtration criteria do not presume any moral principles of the type that are in competition during adjustment of the tripled set: CJs, MPs, and BTs. Through the use of both empirically substantiated theories and normative laws, FP can determine credibility without committing vicious circularity. Arbitrariness can be avoided if the filtration criteria are supported by independently established theory (psychology, sociology, evolutionary psychology, cognitive science, etc.), or include basic principles – even

normative principles – that are uncontroversial and highly and widely corroborated. The candidates that I have proffered so far are the rationality criterion, moral relevance, and formally correct conditions, such as factual accuracy and conceptual clarity. With the filtration process filled-out, we can now put it to work in intuition credibility determination. In the next chapter, I will examine moral intuitions with specific attention to social and evolutionary etiologies. As a result of etiological investigation, the credibility of some of our moral judgments will be impugned, and these judgments will need auxiliary substantiation if they are to be included in the set of considered judgments.

Chapter 4: "Etiologies that May Affect Moral Intuitions"

Introduction

In the last chapter, we considered etiologies and error-disposed conditions (EDCs). If an intuition arises under EDCs, then that intuition thereby lacks credibility.[201] One way we might determine whether EDCs obtain in the generation of an intuition is to examine an intuition's etiology – the causal origin of the intuition.[202] In this chapter, I will continue to examine hypothetical examples in heuristic illustration of EDCs and etiologies.[203] The ultimate goal of this chapter, however, is to move beyond hypothetical examples, and demonstrate how empirical data can affect intuition credibility. In particular, I will consider the ways in which social science and evolutionary etiologies can diminish the credibility of certain intuitions.

In the first section of this chapter, I present a few cases of nonmoral intuitions that, though commonplace, are erroneous. I then move from nonmoral to moral intuitions, and delineate three error-disposed conditions (EDCs) that impugn the credibility of moral intuitions.[204] In illustration of how intuition credibility becomes impugned via these three EDCs, I consider varied empirical studies that suggest, I will argue, that certain sets of moral intuitions should be deemed noncredible.

---

[201] The content of that intuition may be deemed "proxy-credible," however, if auxiliary justification can be found via credible considered judgments or moral principles. This will be discussed in detail later in the chapter.

[202] Joyce (2006) employs a similar concept, which he terms the "genealogy" of an intuition. Etiologies (genealogies) are a subset of error-disposed conditions, as mentioned: for example, disgust is an error-disposed condition, but we wouldn't call it an etiology, in the sense I'm using the term.

[203] One of primary examples centered on the militant abolitionist – an example presented by Nicholas Sturgeon (1992, pp. 97-101). Sturgeon characterizes the militant abolitionist as an individual who has the intuition that slavery is immoral, where this intuition is subconsciously based upon class envy rather than the features of slavery. Given that the causal origin of the intuition is class-envy rather than any intrinsic features of slavery, then the intuition lacks credibility.

[204] These three delineated error-disposed conditions are selections from the list of error-disposed conditions presented in chapter three, and bear upon the later empirical examples I take up later in this chapter.

Erroneous Intuitions

One primary interest of this chapter concerns intuitions which appear cross-culturally. Cross-cultural intuitions are of particular interest to MRE, as well as to other intuition-based methodologies, as they are typically regarded as one way of substantiating the credibility of intuitions. For instance, Norman Daniels (1979a) cites intuitions shared across cultures as representing possible evidence for objectivity. In chapter two, I argued that if a considered judgment is shared across cultures, this considered judgment should be attributed increased credibility: namely, broad credibility. Broad credibility does not guarantee credibility (or objectivity), but it does reduce sources of error.[205]

We should be careful of attributing too much credibility to intuitions – whether moral or nonmoral – on the basis of universality. Universality could be merely indicative of a ubiquitous error-disposed condition, whether biological or cultural in nature. If an EDC is pervasive throughout cultures, then the resulting intuition – even though widely shared – will lack credibility, and its universality will not be corroboration but merely an indication of the pervasiveness of the EDC. In illustration of widespread yet erroneous intuitions, consider a few cases of nonmoral intuitions in the fields of physics, mathematics, and statistical reasoning.

In regard to how physical bodies behave in space, research indicates that most of us have fallacious intuitions. In one study, subjects were presented with a spiral tubing with a ball inside it. They were told that the ball would proceed from the center, spiraling outward toward the opening. Subjects tended to predict that the ball would exhibit a curved, rather than straight, motion once it left the spiral tubing.[206] As Newtonian Mechanics predicts, after the ball exits the tubing, it will proceed in a straight line; the ball will not exhibit curvature motion.

There are several other examples regarding our intuitions of physics: the path a pendulum will take as it falls to the ground after stopping at various points in its swing;

[205] Error that might arise, for instance, from societal or cultural bias.
[206] This experiment has been conducted on children and adults, as well as cross-culturally.

which of two objects, significantly different in size and weight, will hit the ground first if dropped from the same height; the trajectory an object dropped from a moving plane will take. Interestingly, not just adults, but infants as well, were subjects of such studies. These studies suggest that infants share many of the same physics-related intuitions that adults tend to have.

Our erroneous intuitions also extend to statistical reasoning. One infamous fallacy, known as the Gambler's Fallacy, occurs when a person makes bad predictions based on the intuition that past outcomes influence future probabilities. For instance, if a coin-flip results in 10 heads in a row, most subjects will predict that it is more than fifty percent likely that the coin will land "tails" upon the next coin-flip.

The Monty Hall problem illustrates a similar lack probabilistic reasoning. This problem is related as a game-show thought experiment: There are three doors, but a valuable prize only behind one of them, and nothing behind the others. The subject is then asked to then choose one of the three doors. Once the subject chooses one of the three doors, the game-show host holds that door. The game-show host then opens one of remaining two doors, and shows that one door is empty. The subject is then asked if he'd like to change their initial selection from the closed door they've initially chosen, to the remaining closed door they didn't choose. A majority of subjects will decline changing the door they originally selected, often remarking that it seems arbitrary whether or not they change their selection: they presume that there is an equal chance between the two remaining doors. This intuition is fallacious, however: There is in fact a two-thirds chance the unselected remaining door contains the prize, whereas there is only a one-third chance the initially-selected door contains the prize. This answer is so counter-intuitive that many people have a difficult time accepting it as true.[207]

It is unclear why these fallacious intuitions are so pervasive, and to what degree they are error-disposed artifacts due to culture or biology. Such research does provide us reason to pause before we rely too immediately upon our intuitions as truth-tracking or self-evident. Intuitions, even if widespread, can be erroneous.

[207] The problem is easier to understand if the doors are significantly increased, such as 1,000 doors where the subject chooses one door, and 998 doors are opened and revealed as empty. Of the two remaining doors, it seems more intuitively obvious that it is statistically advantageous to switch doors.

In regard to normative intuitions, research in social psychology can provide reasons to question the credibility of some of the everyday intuitions we have. One recent study concerns our intuitions about moral character. According to Epley and Dunning (2000), while an individual has an accurate perception of others' moral dispositions, that individual tends to overestimate his own moral dispositions: "Researchers have demonstrated that people on average tend to think they are more charitable, cooperative, considerate, fair, kind, loyal, and sincere than the typical person but less belligerent, deceitful, gullible, lazy, impolite, mean and unethical – just to name a few" (p. 861).[208] The research by Epley and Dunning substantiate that while subjects could accurately assess these moral dispositions in others, they significantly overestimated these dispositions when in regard to themselves.[209] These moral dispositions are not limited to intuitions of moral character, but how we would act in certain moral situations.[210] These intuitions are not exactly prescriptive moral intuitions, but rather intuitive assessments of moral facts: for example, how likely a stranger is to donate to charity. These still might be considered moral intuitions, however, as they are intuitive moral assessments of the moral character of others and one's own self, such as fairness, kindness, loyalty, deceitfulness, belligerence, meanness, and unethical-ness. Of particular interest concerns how well we deceive ourselves, believing we are far more virtuous than we actually are. Again, these intuitions are not morally prescriptive in themselves, but rather are moral evaluations of individuals. These intuitive evaluations, nevertheless, may *lead* to moral prescriptions, such as, for example, "When divvying up free resources, all things being equal, I should be allocated a bit more than others, as I am more deserving than others."

All of these experiments I've been considering, above, test subject intuitions against substantiated facts. If a subject has the intuition that she is extremely charitable,

---

[208] Epley and Dunning cite the following: Alicke, 1985; Allison, Messick, & Goethals, 1989; Dunning, Meyerowitz, & Holzberg, 1989; Goethals, Messick, & Allison, 1991.

[209] Epley and Dunning (2000) do seem to be assuming a close correlation between virtues/vices and actions. I am assuming that there is a close correlation, but acknowledge that it is a matter open to further discussion and examination.

[210] One study Epley and Dunning (2000) conduct regards how many flowers students expect to purchase for charity during UCLA's "Daisy Days," and how many flowers they do in fact purchase. In addition to this example, they also ask students to forecast their behavior were they to participate in the Milgram's experiment, or the prisoner's dilemma; for obvious ethical reasons, students only took part in the former experiment.

this can be corroborated with whether or not she is as charitable in future action as she is in her predictive perception. The corroboration of nonmoral facts, and even moral character, is easier than in the corroboration of prescriptive moral intuitions.[211] There are no clear and evident "facts" that we can easily get at in the case of prescriptive moral intuitions; rather we talk in terms of credibility rather than truth. The credibility of a moral intuition may be impugned if any of the following three EDCs is the case:

(1) The intuition appears to be based upon no morally relevant features[212]

(2) The intuition is based upon some features that are not morally relevant[213]

(3) Arbitrary factors significantly influence the moral intuition[214]

Keeping these three EDCs in mind, I will examine moral intuitions regarding five topics: incest, kin bias, doing versus allowing, trust, and retribution. The basic argument concerning all of these topics can be represented as the following: If the generation of a moral intuition is sufficiently explained by an etiology that is bereft of any morally relevant features, and after a diligent search no other etiology is discovered that provides MRFs, then that intuition is noncredible.[215]

Incest Taboos

[211] The difference between fact-referential intuitions and moral-referential intuitions is provided some explication in chapter 2. Intuitions regarding factual truths are easily verifiable, at least in simple cases; intuitions regarding moral truths, on the other hand, are less easy to verify, and must be substantiated, in part, via methods of coherence. For this reason, Norman Daniels (1996c, p. 33) characterizes moral truths as dissimilar to observational reports, but more similar to experimental results in science that, in order to be corroborated as factually true, presume the truth of several theories at the outset.

[212] Such as condemnatory moral intuitions in regard to an incestuous sex act when no morally relevant features seem to be present upon which to predicate this intuition.

[213] The degree to which the moral intuition predicates upon morally irrelevant features rather than morally relevant features is the degree to which the intuition is noncredible. For instance, an individual might intuit that killing an innocent child is wrong because children are sentient, but think the killing is *more* wrong because the child has a cute button-nose. I daresay that this feature, a child having a cute button-nose (henceforth, CB-N), is not a morally relevant feature.

[214] For example, the artificial introduction of oxytocin into a person's nasal passage, which results in the subject's generating stronger trust intuitions toward a stranger.

[215] The intuition should be deemed noncredible at least until a point at which an etiology is identified that provides morally relevant features. In the very least, the burden of proof is shifted upon those who claim the particular intuition should be taken as credible. These individuals must present an etiology which provides morally relevant features.

The moral intuition that incest is immoral is an intuition that seems to be widely shared among cultures.[216] Nevertheless, moral judgments concerning certain cases of incest appear to be an instance of the first EDC, listed above, where the intuition seems to be based upon no morally relevant features. Recent studies in moral psychology show that subjects intuit that incest is categorically wrong, and yet when the subjects search for reasons in support of their moral intuition in certain challenging cases, they find themselves "morally dumbfounded."[217]

Jonathan Haidt, for example, conducted an experiment where subjects were presented with the following hypothetical scenario:

> "Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep the night as a special secret, which makes them feel even closer to each other. What do you think about that; was it OK for them to make love?" (2001, p. 814).

Reportedly, most subjects who hear the Julie and Mark story immediately morally judge that the characters' decision to make love was wrong. When asked *why* the incestuous act in the story is wrong, subjects struggled to produce justifying reasons. In the example, there is a negligible chance of pregnancy; both individuals are adults who seem to be in a psychologically healthy frame of mind; no one besides Julie and Mark knows of the act. In essence, nothing ostensibly "wrong" has been done in the scenario. There are no bad consequences – in fact there are good consequences: the act is a pleasurable one which brings Mark and Julie closer together. Furthermore, no autonomy or rights have been violated, and the action doesn't necessarily seem indicative or

[216] The term "immoral" signifies that incest is more than merely distasteful or disgusting, but is thought to be objectively wrong rather than a subjective matter of taste or social convention. Also, the moral intuition that incest is immoral is one shared across cultures and history. This is not to say that in some cases, incest was socially practiced. But a few exceptions do not erase the phenomenon that is a universal incest taboo.

[217] This is a term the studies' authors use, describing the subjects' confusion where they hold onto their moral intuition and yet are at a perplexed loss for substantiation via reasons. Oddly enough, they often laugh and find it humorous that they are unable to substantiate their moral intuition.

encouraging of a bad character.[218] Mark and Julie are not going to engage in sexual relations again. When asked about their moral intuition that condemned incest in this case, the experimental subjects conceded there were no negative consequences or harm being done, but maintained their moral condemnation nonetheless.

No morally relevant facts are evident upon which to ground a condemnation of incest in this case. Nonetheless, our moral condemnation of it – which extends beyond mere personal distaste – sustains. What may we conclude from the presence of the incest condemnation intuition (hereafter, ICI) and the absence of justifying reasons (or morally relevant features) upon which this intuition can be predicated?

It seems reasonable to conclude that the presence of ICI, if morally relevant features are indeed absent, suggests moral intuitions might arise, at least upon occasion, for reasons other than moral truth.[219] In the incest case, we might speculate that a moral intuition that condemns incest might have arisen as an evolutionary product. A sufficient etiology accounting for ICI is evolutionary: ICI arose on the basis of some biological imperative, namely not to waste resources by birthing unfit offspring.[220]

Philip Kitcher, a notable philosopher of science and occasional critic of sociobiology, states, "One of the most defensible claims of human sociobiology is the thesis that human incest-avoidance has the function of lowering the risks of producing defective offspring" (1986, p. 71).

While such reasons to avoid incest are consequential in nature, the condemnation of it is not consequence-referential: that is, the ICI remains even if all the negative consequences are removed. While incest behavior would result in reduced fitness in the past, in a society where birth control is widely and reliably practiced, the detrimental

[218] A virtue ethicist might be able to make an argument that incestuous sexual relationships were bad based on naturalistic considerations and how flourishing qua human beings might be inhibited, which would not seem necessarily to beg the question. However, this is a specific theoretical determination that, I imagine, is not likely to inform the largely pretheoretical moral intuitions of experimental subjects about incest in this case.

[219] "Truth" is intended here as a correspondence relation, where the intuition "X is immoral" is true only if X is actually immoral. In the case of incest, the intuition "incest is immoral" would not correspond to the truth of that intuition, that incest is actually immoral; in fact, the proposition "incest is immoral" would be false.

[220] Another sufficient etiology might be social in nature, yet as we have already seen as we searched for morally relevant features upon which to predicate ICI, no MRFs seem forthcoming.

consequences of incestuous intercourse would practically be eliminated. In this way, we might assert that ICI was predicated on MRFs in the past – namely, upon negative consequences and harms resultant from incestuous relations – but that in today's contemporary, industrial societies, ICI is not adequately predicated upon any MRFs, and so is merely a vestigial moral intuition. Given this verdict, our incest condemnation intuitions should no longer be attributed credibility – in the very least, the burden of proof is shifted onto those who claim such intuitions to be credible to provide a supportive account.[221]

Incest condemnation intuitions are not credible if they are not presently predicated upon any evident MRFs; nonetheless, ICI still arises even in absence of any MRFs. To better understand ICI credibility, we can investigate its likely origin.

In the ancestral environment, ICI would have been based ultimately on its detriment to individual genetic propagation. Genetic self-interest does not constitute an MRF, itself. Nevertheless, if there were a proximate cause which did constitute an MRF, and that MRF was a significant proximate cause of the intuition, then our evolutionary ancestors' intuition condemning incest would be credible. They needn't be cognizant of the MRF, just as long as there was an MRF that was responsible for the intuition, where a relevant counterfactual would obtain in the appropriate way.[222]

For instance, presume that incest had a high tendency to result in harm to the mother and offspring (say given complications in pregnancy and thereafter). We can assert that this harm is a morally relevant feature. Even if the ancestors were not cognizant that incestuous sexual relations were bad because of such harm, its detrimental effects would still constitute an MRF. The counterfactual would hold that if incest were *not* in fact harmful, the evolutionary ancestors wouldn't feel condemnation for it.

---

[221] A virtue ethics response could relate to "natural" human relationships, and how humans have specific modes of flourishing qua human beings, and that our moral psychology is cognizant of this, and so our condemnation of incest is actually our condemnation of the practice being very antithetical to Julie and Mark's character/flourishing and could detrimentally affect those they associate with, given this underdeveloped or badly-disposed character. This would provide an MRF that could justify the credibility of the incest condemnation intuition. One needs to tread carefully when deeming what is "natural" and conducive to flourishing, as the same template as promoted above could also be applied – and seems to have been by some religious leaders – to the condemnation of homosexual relationships.

[222] The counterfactual would be that if the MRF was absent, and other sufficient causes were removed (as the ICI could be overdetermined), then the incest condemnation intuition (ICI) would also be absent.

However, incestuous sexual relations were in fact proximately harmful, which in turn was ultimately bad for individual genetic propagation.

In this case, ICI would arise from the effect of incest on both genetic self-interest, as an ultimate cause, and harms, as a proximate cause.

If incest were not against an individual's genetic self-interest, as the ultimate cause, then there would be no incest condemnation intuition. It could be the case that ICI were determined by some other ultimate cause, such as hypnosis or neurophysiological tinkering (as presented in previously considered thought-experiments presented in chapter 3). Similar to the above case regarding proximate causes, these *possible* alternatives are irrelevant. The counterfactual test pertains to the actual causal chain, not a hypothetical causal chain. In our case, the causal chain includes the proximate cause of harms resultant from incest, and genetic self interest, as the ultimate cause – both of which can be said to determine ICI. The proximate cause, in this case, provides an MRF, whereas the ultimate cause does not. ICI can be deemed credible, then, as it can be predicated on resultant harms, which represent an MRF.[223]

To illuminate the point, discussed above, between proximate and ultimate causes, it might help to consider another example. Consider the following case: Dan morally condemns Mike because Mike killed Dan's son. Dan morally condemns Mike, at least in part, for the suffering Mike has caused via his action, given that Dan loved his son and grieves the loss.[224] The *ultimate* reason Dan morally condemns Mike, however, is not love itself, which is only a proximate cause, but because of genetic self-interest, which has evolutionarily developed this emotional system in the human species as a constructive reaction that optimizes individual gene propagation. The proximate cause, however, is (parental) love. The (parental) love, and the suffering caused as a result of it given this loss, seems to be a morally relevant feature upon which the moral condemnation can be appropriately predicated. We could imagine eliminating all love as

[223] By "discordant" I imagine something close to a Hobbesian State of Nature: "solitary, poor, nasty, brutish and short." These bleak adjectives can be cashed out in terms that seem morally relevant: where one's life is lonely, painful from destitution, rife with fear, brutal in its inflictions, and dismal in its preclusion from possibilities of flourishing which can only blossom in the soil of longevity.

[224] Dan can morally condemn Kyle not merely because Kyle has caused *his* suffering, but because Kyle has caused suffering in general.

a proximate cause, as well as eliminating all other proximate causes, and just say that Dan morally condemns Mike, simply because Mike's killing of his son reduced the propagation ratio of his genetic material. However, appealing to the propagation of genetic material would not be a morally relevant feature, in and of itself, upon which a credible intuition could be based. Thereby, the intuition would be noncredible. However, this does not show that the original case to be noncredible. It is, in fact, credible upon (parental) love, which is the basis of the suffering Mike has caused to Dan by killing his son. If the ultimate cause – genetic self-interest – is the only candidate to provide any morally relevant features, but fails to do so, then the moral intuition fails to be credible.

The ultimate aim of this chapter is to establish how empirical research can be relevant in determining the credibility of intuitions. As we see in regard to incest, one way a moral intuition can be deemed noncredible is when an etiology, in this case an evolutionary etiology, provides a sufficient explanation for that intuition where morally relevant features in the case appear to be lacking. In certain cases of incest, no other morally relevant etiology seems evident by which to vindicate the credibility of morally intuitions condemning the practice.

I will postpone, for now, an examination of empirical research corroborative of evolutionary etiologies, and will continue the examination of evolutionary etiologies by presenting two hypothetical experiments as a heuristic template. After establishing the way in which social psychological findings could impugn intuition credibility, I will move from the social psychology to evolutionary biology and evolutionary psychology, and consider intuitions relating to kin preference in light of sociobiology research. If empirical evidence suggests the presence of a psychological or biological bias toward one's kin, this evidence would represent an error-disposed etiology, which would commensurately diminish the credibility of certain kin-related intuitions.

Bias toward Appearance Similarity

Imagine the following thought experiment. John is asked to arbitrate, as a judge, over two cases. He is to base his judgments entirely upon his moral intuitions, and refrain

from appeal to any moral principles he may hold. The participants involved – Frank and Tim – are both strangers to John. We can imagine a series of cases involving some minor offense: for example, the case of a fender-bender accident between two parties. John, as the judge, has to decide if person A, the collider, is at fault or person B, the rear-ended, is at fault. If John determines person A is at fault, person A will have to pay $100 to person B. If person B is judged to be at fault, person B will have to pay $100 to person A. Imagine that the hypothesis being tested is whether or not physical appearance – namely physical similarity to the subject – has any undue affect on the subject's moral intuitions in regard to others. In our example, imagine that Frank has a similar appearance to John, whereas Tim does not.

In case 1, Frank is in person A and Tim is in person B. After hearing their testimony, John generates the intuition that person B, occupied by Tim, is at fault, and he awards person A, occupied by Frank.

In case 2, the identical case, Frank and Tim switch positions, where Tim is now A and Frank is now B. John hears the testimony again, testimony which is identical in both word and nuance, and yet not delivered from the same individual. Conveniently for our hypothetical example, John has conveniently been memory-wiped of any recollection of the previous case 1. After hearing the testimony and surveying the facts, identical to case 1, John generates the intuition, contrary to case 1, that position A is at fault, and so awards position B, which, again, is Frank.

In each case, before the memory wipes, John reports that he made his decision on the basis of his moral judgment regarding who was morally at fault.

Imagine that this experiment is repeated several times, with John as our subject. The participants, Frank and Tim, are replaced with other individuals: always one participant who was significantly similar to John in appearance. Also, the arbitration scenarios were changed significantly, running the gamut of cases that typically confront judges in small claims court over their careers.[225]

[225] The focus is upon relatively minor scenarios rather than significant ones as it seems more plausible that bias would insinuate itself when there were not salient countervailing conditions. For instance, if the case was one of murdering the other's family member, we might find it hard to imagine John's bias extending *that* far. Of course, maybe that is too quick. We can imagine that court cases arbitrated by racist

In each case in this series of experiments, John always generates moral intuitions in favor of the participant similar to him in appearance. From this, we could reasonably conclude that John was not generating intuitions under the relevant cognitive conditions, but was generating moral intuitions under an EDC of bias: namely, bias toward those of similar appearance to him. The credibility of the John's intuitions, at least in cases involving these elements, would be significantly diminished. The reason for this diminished credibility is that the features from which John is determining moral fault are features that were not causally responsible for the accident itself: that is, the feature of similar appearance, in this case, is not a morally relevant feature upon which blameworthiness can predicated.

If this case were not particular only to John, but turns out to be a phenomenon exhibited in all experimental subjects, we would have the same reasons to believe each and all of these subjects to have noncredible intuitions in such cases – cases where someone in the experiment shares a resemblance to the subject who was arbitrating distribution based on desert. Furthermore, even though this category of moral intuition – that individuals similar in appearance seem to be more morally deserving, as intuited by the acting judge – would be ubiquitous and cross-cultural, this would merely indicate this EDC was pervasive, not that this intuition was credible.[226]

Bias toward Genotypic Similarity

---

or sexist judges often pivot upon who sits with the prosecution and who with the defense. We can imagine that many of these judges base their decisions in large part on biased moral intuitions rather than just the facts and morally relevant features of the case. The judges themselves might concur that race/sex should not be taken into consideration, as the law presumes equality, and seek to remain faithful to this; nevertheless, their moral intuitions might be polluted by this underlying bias, precluding any attempt at objectivity.

[226] Though there are no direct empirical studies that substantiate this hypothetical case directly, some more tangential cases might provide some reason to believe in bias based upon appearance. One sociological study suggests that we are more likely to perform supererogatory acts toward a particularly attractive person, even if we are never likely to meet this person: in particular, the study recorded the percentage of times a lost wallet would be returned to an owner if the picture inside of the wallet-owner was of an very attractive individual versus an average-looking individual. Psychology studies of beauty, asking survey respondents to take a look at various pictures and describe the character of the subject depicted: "respondents tended to describe very attractive individuals with more positive characteristics than average-looking individuals. People with more attractive faces were assessed to be more successful, contended, pleasant, intelligent, sociable, exciting, creative and diligent than people with less attractive faces" (Braun, C., Gruendl, M., Marberger, C. & Scherber, C., 2001).

Consider a slightly altered version of our fender-bender hypothetical experiment. The scenario is the same, except that rather than being similar to John in appearance, one of the subjects shares significant genetic material with John. While the two know they both are related to one another, they have had no social association with each other and were raised in two different environments from birth. Also, to distinguish this from the previous version, let us assume that there is no significant physical similarity of appearance to John; in fact, it's possible that the non-genetically-related subject could look more similar to John than the relative.

Imagine that, in this fender-bender arbitration scenario, Frank is John's brother and Tim is not (closely) genetically related to John. John once again arbitrates on the basis of his moral intuitions. It turns out that irrespective of whichever position Frank occupies, John always generates the moral intuition that sides with Frank.[227]

If the experiments were extensive enough in number, and wide enough in scope, for instance where Frank is switched out by other genetic relatives socially unknown to John, we might conclude that John, though he has no socially cultivated filial feeling or association with his genetic relatives, that John's moral intuitions are still biased given genetic similarity. We should also stipulate that the experiments show that John did not prefer those who looked similar to him in appearance, but would prefer genetic relatives even if they looked less similar to him than the other participant.

If these results were established, it would be reasonable to conclude that John was biased on the basis of genetic similarity to the subject. We could then deem situations to be error-disposed if the situation involved close genetic relatives of John. If John did have moral intuitions in these situations, we would have reason to question the credibility of his moral intuitions.[228] The reasons John's moral intuitions lack credibility in this case

[227] Chagnon and Bugos (1979) provide evidence that knowledge of genetic relatedness of the participants would greatly improve one's chances of predicting who would take sides with whom in a complex Yanomamo physical conflict.

[228] Unless, perhaps, he were to go *against* his disposition and award the non-genetically-related person – in which case we'd think his intuition must be correct, since it arose in spite of this contrary bias. Similarly, we might find the ruling of a racist judge credible if he awarding damages to a Black man. It would not be a credible ruling merely because we happened to agree with it, but that his bias pointed in the other

is similar to the previous case: namely, the features from which John is determining moral fault are features that were not causally responsible for the accident itself: that is, the feature of genetic similarity, in this case, is not a morally relevant feature upon which blameworthiness can predicated. This revised case illustrates how bias could, at least theoretically, be biologically-based. Empirical research can support the hypothesis of biological bias. For instance, one interesting empirical study by Segal and Hershberger shows that monozygotic twins are significantly more likely to enter into cooperative exchanges with each other (when playing prisoner dilemma type games) than dizygotic twins (1999, pp. 29-51).

John's relationship to his relatives is only genetic; he is socially alienated from the genetically-related subjects in the experiment. Our relationship with genetic relatives is typically different: We benefit our family members largely because of our feelings of association with them or our conscious sense of moral obligation given cultural and societal structures and institutions (which presumably isn't true for John). What's at issue is whether our feeling and sense of obligation is biased.[229] If evolutionary psychology provides good reason to suspect our moral intuitions are biased toward genetic relatives due to a *biological* imperative, and we can find no morally relevant features upon which to predicate this preference, then we have good reason to suspect that these intuitions are not as credible as initially thought. Sociobiologist, John Alcock, explains:

> "An awareness of the ultimate reasons for our eagerness to make moral judgments and the realization that our emotions really work on behalf of our genes ought to make us less self-indulgent about our feelings, perhaps encouraging us to be a little more cautious on the moralizing front, a little more reluctant to express moral certitudes, a little more introspective, a little less likely to assume that whatever feels right to us is good for something other than our genes." (2003, p. 206)

To determine whether we, as human beings, may be unduly biased toward kin, we need to investigate the primary mechanism of evolution in reference to our kin: namely, kin selection.

---

direction.

[229] I accept the possibility that it is morally justifiable to favor individuals with whom we are in social relationships. Bias is when such justification is outstrips its morally relevant features. For instance, a mother might have moral reason to provide for the needs of her child over the (greater) needs of another child; nevertheless, she might not be justified in providing for the mere *wants* of her child over the dire needs of another child.

Kin Selection

Kin selection, a well-established mechanism of evolution, is the disposition of an individual to make certain sacrifices to benefit its relatives. Since relatives share genetic material with the individual, it is the genetic interest of the individual to make limited sacrifices in order to ensure the survival of its genes in relatives. Stated non-teleologically, individuals who make sacrifices that benefit their relatives, thereby promoting the survival of these relatives, tend to increase their individual genetic propagation to a greater degree than both of the following:

(1) Individuals who do *not* make such sacrifices, which is subsequently deleterious to the indirect propagation of genetic material via relatives

(2) Individuals who make sacrifices to individuals, irrespective of an animal's genetic similarity

Alcock explains how the disposition towards benefiting kin has evolved:

> "In the past, individuals who tended to act in concert with their relatives would have sometimes indirectly propagated the genes they shared in common with each other. In contrast, those who regularly harmed their relatives' reproductive chances would sometimes have reduced the number of those shared genes passed onto the next generation. Over time, this process should eliminate any distinctive genes that contributed to the development of personalities invariably indifferent or hostile to one's relatives" (2003, p. 200).

Darwin was the first biologist to understand kin selection. At first, he was puzzled how social insect colonies could have evolved in nature. Only later was he able to arrive at a logical solution: as long as sterile workers promoted the survival and propagation of other genetic family members, the workers' hereditary attributes would be carried into the next generation.[230] This explanation reconciles the self-sacrificing behavior of worker bees, who sting honey-robbing animals, with natural selection. Though this self-sacrificing behavior is immediately detrimental to the individual worker bee, genetically it is advantageous to the individual worker bee's genes.

[230] That Darwin arrived at this solution is particularly impressive, since he had no knowledge of Mendelian genetics.

In humans, the most obvious example of kin selection is parental care. Offspring are the vessels in which genes persist in time and are propagated. Thus, we should expect mothers to take care of their children -- even to if it is to the mothers' individual detriment. In regard to parental valuation of offspring, Hume (1742) writes:

> "Nature has given all animals a like prejudice in favour of their offspring. As soon as the helpless infant sees the light, though in every other eye it appears a despicable and miserable creature, it is regarded by its fond parent with the utmost affection, and is preferred to every other object, however perfect and accomplished. The passion alone, arising from the original structure of human nature, bestows a value on the most insignificant object." (1987, pp. 162-163)

Because of the genetic advantage of benefiting kin, Alcock explains that evolution has selected for the kin selection mechanism in humans:

> "…selection has evidently favored people with the motivation mechanisms, emotional systems, and intellectual capacities that enable us to learn kinship categories, establish kin-based links with others, educate others about genealogical relationships, and feel a sense of solidarity and cooperativeness with those identified as relatives, especially with our close relatives." (2003, p. 201)

Kin selection also includes siblings, as they carry the same amount of related genetic material as a parent or child. Although kin selection also extends to extended family, Alcock stipulates, "But as degrees of relatedness decline, the chance that altruism will be genetically profitable also declines (p. 200). Thus, we would expect to find in animals, a diminishing degree of altruism in proportion to diminishing relatedness. Indeed, this is what we do find. Studying bee-eaters, a species of bird, Stephen Emlen (2001), a prominent biologist, found that daughters-in-law fail to provide food to siblings of their mates, that males with excellent chances of reproducing defiantly resist their fathers attempts to recruit them as helpers at the nest, and that the probability of helping declines in families with replacement mates when the helper would be assisting in the production of half-siblings rather than full siblings (p. 201).

Empirical research also suggests predispositions toward genetic favoritism. An anthropological study of the tribal Yanomamo culture by Chagnon and Bugos reveals that, during times of complex physical conflict within the tribal community, subjects' alliances predictably matched up with the subject's knowledge of the degree of genetic relatedness to one another (1979, pp. 213-238). In regard to those outside their tribal

communities, the Yanomamo were known for thinking nothing of killing foreigners. Also, as previously mentioned, in a recent sociological study by Segal and Herschberger (1999) involving cooperation in prisoner dilemma type games, monozygotic twins were significantly more likely to enter into cooperative exchanges with each other than dizygotic twins.

Dispositional Bias and Moral Intuitions

The studies introduced above corroborate the suggestion that genetic bias does exist and affect our attitudes and behavior. Yet it is a further step to assert that this genetic bias results in biased, and thereby noncredible, moral intuitions. My claim is that such genetic predisposition, such as a dispositional bias toward kin, leads to emotional favoritism: more sympathy toward siblings than strangers, less disgust of one's children than non-related children, less anger toward kin non-reciprocators than unrelated non-reciprocators, and so forth. This "emotional favoritism" leads, in turn, to biased moral judgments. The claim can be represented as the following two-step argument:

1. Genetic predisposition (e.g., for kin bias) leads to emotional favoritism (e.g., more sympathy towards, less disgust by, less anger toward).[231]
2. Emotional favoritism leads to biased normative valuation (e.g., moral).[232]
3. Therefore, genetic predisposition leads to biased normative valuation.

The first premise is to be substantiated by empirical scientific and social scientific research, such as the empirical studies proffered in this chapter. In the same way, the second premise is to be substantiated by empirical research. The empirical evidence may

[231] For example, "The Cinderella Effect" regarding abuse of stepchildren, which I discuss later in the chapter. I presume the reason for this abuse is precipitated by an emotional disposition in the stepparents: for instance, that they are more disposed to becoming angered and less sympathetic toward their stepchildren than their own children. The studies, which I will cite later, assert that the explanation of this disparity in treatment cannot be sufficiently attributed to the salient sociological factors likely to be causally associated with disparity in treatment, such as having associated with one's own child longer than one's stepchild.

[232] As we saw in chapter 3, disgust is an emotion associated with moral valuation. If one is more easily disgusted by the behavior of one's stepchild than one's own child, this might lead to a disparity in moral valuation.

not be conclusive; nevertheless, it does provide some substantiation for the assertion of each premise, and thereby, additively, support of the argument's conclusion. There could also be studies, though I am not currently aware of any to date, which investigate a direct link between genetic predispositions and moral intuitions in regard to kin as opposed to strangers, similar in aim to the hypothetical cases involving John and the fender-bender case. Excluding this direct link, the empirical gap between a genetic predisposition and moral intuitions is one that can be filled by further studies in this area.

Genetic Bias as an Error-Disposed Condition

Evolutionary biology establishes that kin selection is a primary evolutionary mechanism. The question before us is whether kin selection has significantly influenced human beings, and how it has influenced our moral intuitions. To the degree that sociobiological, sociological, psychological and anthropological research suggests that our moral intuitions may be significantly influenced by genetic relatedness, we have grounds to question the credibility of intuitions that are generated when this condition is present.[233] Genetic relatedness, by itself, is not a morally relevant feature, just as eye color, skin color, or physical similarity is not an MRF.[234] Therefore, if genetic relatedness does bias our moral intuitions, we have reason to attribute less credibility to these intuitions.[235]

In the second hypothetical experiment John, we saw that genetic relatedness was biasing his moral intuitions in that he consistently favored the individual who was genetically related to him, irrespective of that individual's position in the scenario.

[233] The connection between evolutionary forces and moral intuitions are of course distant and unclear. Nonetheless, certain empirical studies might substantiate that there is a biological basis for certain moral sentiments that we have. As asserted earlier in the chapter, if cross-cultural findings indicate a pervasive moral condemnation of incest, this would seem to corroborate the hypothesis that we have biologically inculcated predispositions against incest. These predispositions might be suspected to manifest as moral intuitions if, for instance, the moral intuitions turn out not to be grounded in any apparent morally relevant features.

[234] "Genetic relatedness," refers minimally to two or more individuals having similar genetic material, and some knowledge – whether conscious or subconscious – of this similarity; it does not presume family, proximity, or any other relation among these individuals.

[235] This point was illustrated in the fender-bender example above, where John judged blameworthiness in favor of his genetic relative in all cases, irrespective of that relative's position in identical cases.

Unfortunately, the ideal conditions of this hypothetical experiment outstrip the practical limitations of science; commensurately with these limitations, our conclusions will not be as conclusive.[236] Nonetheless, some empirical studies do suggest that individuals are inclined to be biased toward those whom they, consciously or subconsciously, identify as genetically related to them. To the extent that the sciences can empirically substantiate an evolutionary explanation of genetic bias, we will have a proportionate reason to question the credibility of moral intuitions in regard to kin.[237]

Sociobiological research does in fact provide empirical findings which would corroborate the hypothesis of genetic bias in humans. For example, multiple studies by Flinn, Leone, and Quinlan have shown that stepchildren tend to receive significantly less attention and resources from a stepparent than their genetic parent; moreover, this tendency is cross-cultural (1999, pp. 465-479). Beyond stepchildren not just receiving resources, studies have shown that stepchildren are far more likely to be victims of abuse. A study by Martin Daly and Margo Wilson (1987) reported that for every 10,000 children four years or younger in families with a replacement mate, about 120 were victims of child abuse; whereas only three children were victims of abuse in families with both genetic parents present. Restricting their data to fatal child abuse statistics, a child is seventy times more likely to be killed by a stepparent than by a genetic parent.[238] There is, of course, a gap between maltreatment of stepchildren, as opposed to one's genetic children, and moral intuitions that such unequal treatment is justified. Further psychological research would have to be done to establish that the abusers feel morally justified in this unequal treatment, and generate disparate moral intuitions regarding how they morally ought to treat stepchildren and genetic children, respectively.

---

[236] For instance, we usually cannot isolate the genetic relation from the social relationship between two individuals; the social relationship is in fact heavily defined by the genetic relatedness. Also, experiments cannot be repeated on the same subject under the same conditions (unfortunately, there are not such things as tidy "memory wipes;" if there were, it'd make psychological experimentation much easier).

[237] Our intuitions or moral preference for kin may find some auxiliary justification, but this will be from sources other than the appeal to the direct intuition in regard to kin.

[238] The researchers are aware that social factors may contribute to an adult's disparate treatment concerning children and stepchildren. They argue that the data shows that the disparity outstrips the social factors, where merely a sociological explanation would be insufficient to shore up the disparity.

Another study, conducted in Canada (Littlefield and Rushton, 1986), shows a correlation between emotion and reproductive capability in kin: particularly, the intensity of parental grief at the thought of losing a child at various life-stages. In the study, adults were asked to imagine the death of children of various ages and estimate the degree of grief would be experienced by the parents. The prediction, based on kin selection, is that grief should progressively increase toward adolescence, when the child is capable of reproduction, and then begin to drop off. When this data was graphed and compared to the reproductive-potential of a hunter-gatherer people (the !Kung of Africa) -- which most closely approximates our ancestral environment -- the correlation of the grief curve to the reproductive-curve was almost perfect (1986, pp. 797-802).

Given such research that matches up psychological dispositions to the mechanism of kin selection, there seems to be reason to suspect that we might be biased toward those genetically related to us: especially if our emotional predispositions influence our moral valuations, which current cognitive models in experimental moral psychology seem to suggest.[239] If we are indeed genetically disposed toward kin bias, this bias should be considered an EDC, thereby diminishing the credibility of subsequent intuitions. In this way, evolutionary psychology may provide evidentiary reasons for us to not take our intuitions as self-evident: namely, because those intuitions might have arisen, at least in part, as a consequence of genetic bias.[240]

Kin Preference as an Error-Disposed Condition

Considering the possible effects of kin selection upon our moral intuitions, Singer explains that, though we take it as morally right to give priority to our kin and associates over strangers, we should not presume the cross-cultural pervasiveness of this moral conviction is evidential of its truth:

[239] Recent cognitive studies, such as those led by Jonathan Haidt (2001), involve fMRI scans of brain activity during moral evaluation of characters in moral dilemma scenarios. The data from these studies suggest that emotional reaction is prior to rational assessment, which Haidt calls “The Social Intuitionist Model,” which was introduced in chapter one.

[240] As illustrated in the hypothetical example of John and his predisposition toward genetic bias.

> "It might therefore appear to be a moral conviction which, not being the result of any specific cultural prejudices, has some claim to acceptance as a self-evident principle of morality. A biological explanation of the prevalence of kin preference undermines this claim. If the moral conviction that it is right to give priority to our families rather than to strangers derives from the evolutionary process of gene selection, it loses whatever credence it seemed to possess as a self-evident moral truth. It might, of course, still be a desirable way of living; but that is now a question open for debate" (1982, p. 56).

Singer makes several points here. First, he asserts that a biological explanation of kin preference undermines it as "self-evident."[241] It's not quite clear what the argument is that leads to this conclusion. I presume that he is not suggesting kin preference is undermined from the mere fact it is a product of evolution. Surely all of our intuitions are determined either biologically or culturally; this doesn't necessarily imply these origins undermine their credibility. I believe the best interpretation of Singer's assertion is that the biological explanation heavily influences kin preference, and so our "self-evident" intuitions of kin preference may have arisen under error-disposed conditions.[242]

Given the biological explanation of kin preference, Singer challenges the presumption that the universality of certain moral intuitions – the fact they are present in almost all societies – should be accepted as confirmation of the intuition as a self-evident moral truth, or at least an intuition with a high degree of credibility. Universality could be the result of a biological bias produced in the course human evolution, which would thereby tend to be universal among human beings and cultures. Universality of a certain intuition is corroborative, as long as that intuition doesn't arise under conditions of bias, which may tend to be either biologically universal or socially universal.[243] As specified in chapter 2, cross-cultural agreement lends *broad* credibility to considered judgments and principles, as this broad coherence eliminates sources of error: for example, it reduces the probability that initial judgments are a mere artifact of cultural bias.

---

[241] To clarify Singer's point in relation to MWRE, moral convictions – which I presume to be equivalent to intuitions – are not taken to be self-evident moral truths by MWRE: they are not direct apprehensions of moral features of the world, but are only provisional starting points of theory construction, and only enjoy initial credibility.

[242] That is to say, that we mistake these intuitions as moral imperatives when they are *merely* biological imperatives, rather than that these intuitions are not morally normative merely *because* they are biologically determined.

[243] Biological and social forces are likely interconnected; indeed, sociobiology and evolutionary psychology suggest that social behavior/structures are, at least in part, influenced by our biological makeup as human beings

In investigating our moral responsibilities, Peter Singer claims that there is inconsistency in our moral intuitions.[244] He presents the example of a child drowning in the shallows of the lake (1972). You could save the child, but it would require you to sacrifice some of your resources to do so (e.g., expensive shoes). Our intuitions tell us that it would be immoral of us not to make this sacrifice and aid the child. Singer lists some facts as morally irrelevant: the ethnicity/nationality of the child; if there are other people around who could help, but are not helping; etc.[245]

Singer (1972) claims that distance is not a *morally* relevant feature, so if the child is drowning in front of you or starving in a far away country, the two are identical cases in their morally relevant features. He acknowledges that distance is *psychologically* significant, in that we're less emotionally moved in this case than the drowning case, but he states that the differences are not *morally* significant (pp. 230-231). Even if we had the misfortune of happening upon a drowning child every year, each year we would feel morally required to save that child. However, we do not have the same sense of moral obligation to send an annual charitable contribution overseas (unless compelled by circumstances such as religious prescripts, community pressure, etc.). One is seen as morally required, the other as morally supererogatory.[246]

---

[244] This inconsistency presumes that we can effectively help starving children overseas just as easily and reliably as we can save the drowning child. There is good empirical evidence to believe we can in fact effectively help in many, if not all, of the dire situations around the world in which children are suffering from lack of food, shelter and medical care. Also, I am presuming Singer's "weak" or "moderate" version, where one need only sacrifice resources that are luxuries, which hold no "moral significance" in the robust sense that it gives life meaning: e.g., providing necessities for our children, neighbors, community.

[245] Regarding this last factor, it seems that we feel less morally responsible to assist an individual if others are also present who can help, but are not. This is a termed "diffusion of responsibility." In a study by Darley and Latane (1968), a subject was placed alone in a room where he is told he can communicate with other subjects via intercom. During the experiment, one other "subject" pretends he is having a seizure. The study found that the duration the subject waits to inform the experimenter of the seizure varied inversely with the number of other "subjects"; in some cases, the subjects never informed the experimenter. I put "subjects" in quotes because there were no additional subjects in the experiment, but these were merely voice recordings of actors who were playing experimental subjects. This experiment was the first experiment of its kind to consider diffusion of responsibility or "bystander apathy," and was inspired by a real-life case involving Kitty Genovese, who was stabbed to death by a serial rapist and murderer over the course of 30 minutes, all within view of purportedly 38 witnesses who failed to help the victim or call the police in a timely manner.

[246] This topic is explored further in relation to Unger's (1996) reprisal of Singer's drowning child example: A bleeding man on the side of the road needs a ride or else he'll die; however, giving him a ride will ruin your $200 leather car seats. This is contrasted to receiving a letter in mail, requesting a $200

An infamous psychological experiment, The Milgram Study, shows psychological effects upon our intuitions regarding moral responsibility.[247] The Milgram Study was conducted before more stringent ethical standards were established in experimental psychology. In the experiment, the "teacher" subject is instructed by an authority-figure, a scientist, to electrically shock the "learner" subject every time the learner incorrectly answers a question on a memory test the teacher is orally giving him.[248] There are several conclusions that can be inferred from this experiment: (1) Authority figures, who issue orders and claim responsibility, makes us feel we are significantly absolved of moral responsibility (2) Personal association with an individual makes us less likely to subsequently inflict harm upon the individual (3) Proximity matters to us emotionally and influences our ethical judgments.[249]

Research in psychology and sociology reveal when trivial factors have significant effects upon our moral intuitions and valuations. Such studies show that intuitions are affected by several nonmoral factors. For this reason, it might be difficult to determine to the extent moral intuitions should be deemed credible. However, this task is not insurmountable. Research in relevant sciences – such as moral anthropology, psychology, cognitive neuroscience, evolutionary biology, sociology, and so forth – can provide

---

dollar donation. Greene (2003) finds a similar "out of sight, out of mind" correlation in the experimental subjects consider this set of cases as they did in the previous trolley case: namely, proximity relates to emotional salience which in turn affects a subject's moral intuitions.

[247] The inferential conclusions of what the Milgram experiment in fact shows is a matter of some controversy. The experimental results are often taken as indication that institutional structures have an effect on individual behavior, in which the individual submits themselves to be a cog in an institutional machine, where they no longer view themselves as a moral agent. I presume that their subjugation of themselves as a cog to the greater machine first presumes that they consider it to be, overall, morally legitimate: for instance, even if the pain inflicted upon the learner is a moral harm they are inflicting, they believe it must be for the greater good of scientific knowledge. For the purposes of this dissertation project, however, I do not wish to rely too heavily upon this experimental example, as its conclusions, while interesting regarding moral agency, are somewhat oblique.

[248] The learner subject is not actually a subject, but is actually an actor playing an experimental subject, and is not actually receiving electric shocks, though a prerecorded tape convinces the teacher that the learner is being shocked with progressively increasing voltage.

[249] The degree to which the "teacher" obeyed the instructions of the authority figure had the following correlation: It was proportional to the perceived authority of the figure (the way he was dressed, the academic prestige of the institution he was representing); it was inversely proportional to the degree to which the "teacher" had associated with the "learner" before the experiment; it was inversely proportional to the proximity between the "teacher" and "learner."

evidence that can illuminate our moral judgments and can assist in determining credibility.

Evolutionary etiologies of intuitions diminish credibility only in those cases where there is significant bias and unawareness or self-deception about this bias. If there is good reason to believe the reason we have a moral conviction is *only* due to evolutionary inculcation, then the credibility of the intuition might be undermined (and as such, it would be excluded from the set of considered judgments). The burden of proof is thereby shifted onto the individual claiming kin preference is a justified moral prescription; the individual must then find justification for this moral conviction.

This moral conviction can be redeemed from other avenues: for example, by supporting it via moral principles. For instance, preferential treatment to kin, rather than being justified "bottom-up" – from the preliminary credibility of considered judgments – it could be justified from a "top-down" approach: on the basis of rule utilitarianism, virtue ethics, feminist ethics, and so forth. If preferential treatment of kin is justified in this way, rather than an instance of a credible considered judgment, it will instead have the status of a corollary derivable from a moral principle, where the moral principle has been attributed systematized credibility on the basis of their coherence with considered judgments, moral principles, and background theories. In this way, whenever an individual has a moral intuition that she can give preference to a relative rather than a stranger, this intuition will not, itself, be credible. However, practically speaking, we may treat this moral intuition as "proxy-credible," presuming the prescriptive content of the intuition is found to be identical to the content of a moral principle that has systematized credibility after adjustment via MWRE. We should be careful not to mistake this moral intuition, an initial judgment, as a considered judgment, which enjoys preliminary credibility, and which can be used to construct and test moral principles.[250] A second way a noncredible intuition can be allocated proxy-credibility is if it is identical in content with considered judgments. I discuss these concepts further in the next section.

[250] Though one moral principle, if substantiated via significant coherence, can serve as a test of another moral principle in the same way one scientific hypothesis, if substantiated, can serve as a test of another scientific hypothesis: e.g., If the hypothesis is that striking a match creates more energy, whether potential or kinetic, in closed system, this hypothesis would be rejected on the basis that the hypothesis violates the law of conservation of energy.

Auxiliary Justification and Proxy-Credibility

A moral intuition may suffer undermined credibility in light of scientific research. The moral intuition can still be attributed "proxy-credibility," however, if this intuition is provided "auxiliary justification," on the basis of either considered judgments or moral principles.

Proxy-credibility means that the content of the noncredible intuition is identical to the content of considered judgments or is derivable from moral principles, which enjoy preliminary and systematized credibility, respectively. That is, though the initial judgment is not itself credible, via bottom-up justification as a considered judgment arising under relevant cognitive conditions, nor as a moral principle gaining credibility via coherent systemization in MWRE, the noncredible initial judgment is proxy-credible, as its content is identical with another considered judgment, or it is derivable as a kind of corollary from a moral principle.

In illustration, consider the militant abolitionist's intuition that slavery is wrong. Though this intuition, itself, lacks credibility as it is not based upon any morally relevant feature (but is rather resultant from the abolitionist's class envy), the militant abolitionist's intuition "slavery is wrong" could still be attributed proxy-credible in one of two ways. First, the intuition could gain proxy-credibility from the fact it is content-identical to one or more moral intuitions that are indeed determined to be credible as considered judgments. The sympathetic abolitionist, for instance, has the identical intuition that slavery is wrong, yet her intuition is credible as it is based upon an intrinsic morally relevant feature of slavery, such as the infliction of suffering or violation of autonomy. The second way proxy-credibility can be attributed is if the content of the noncredible intuition could be derivable, as a corollary, from a substantiated moral principle. For instance, if coerced human subjugation is found to be a practice that cannot be consistently willed as a universal law of humanity, then the determination that American slavery is morally impermissible is a derivable corollary from that principle.

For this reason, we might not disabuse the militant abolitionist of the content of his intuition, though we might try to change its foundations.[251]

Similarly, to take another example, we might countenance an entrenched divine command theorist for his morally condemnatory intuition against murder – justified on the basis of Zeus' prohibitive commandments – rather than disabusing him of his intuition as it is noncredible due to its basis upon both a non-existent entity and problematic normative predication upon divine commands. We might find his intuition to be noncredible, but we might not make muck of it.[252] Proxy-credibility can be allocated to his intuition on an auxiliary basis: for instance, to the extent that the intuition conforms to identical intuitions which *are* substantiated via preliminary credibility; or to the extent that the intuition is derivable from moral principles which are substantiated via systematized credibility. Proxy-credibility should only be allocated to an intuition to the degree that it is substantiated from a moral principle or is content-identical to a considered judgment, with systematized credibility or preliminary credibility, respectively.

---

[251] The concept of "justified true belief" appears to work in a similar way to our notion of proxy-credibility. A person can believe that proposition is true and be justified in that belief; nevertheless, that proposition might be true for the reasons other than those the person presumes. Nevertheless, though the person believes the proposition is true for reasons that do not causally make it the case, the proposition still is in fact true and his belief is therefore true, though this belief might not be considered knowledge since it just "happens" to be correct. In the case of initial moral judgments, the belief that a moral prescription is true might not be upon a justified basis (i.e., MRFs); nevertheless, if the noncredible initial judgment is identical in content to credible CJs or MPs, it is still proxy-credible.

[252] One question to consider is if we feel the same way about the militant abolitionist, who has the noncredible initial judgment that slavery is wrong (arising from his own self-interest rather than intrinsic features of slavery). Like the example of the mother and her kin preference intuition, this is a practical matter that needs no resolution in this project. It is an interesting consideration, nevertheless. Presumably, we would first want to reroute the abolitionist's intuition to be predicated upon morally relevant features, such as the suffering of slaves. If his moral intuition could not be rewired in such a way, his intuition would still be allocated proxy-credibility since it is identical in content with credible considered judgments and/or moral principles that slavery is wrong. Whether or not an intuition should be given amnesty, when it happens to be proxy-credible on an auxiliary basis, is to be determined by the practical implications of allowing a noncredible intuition to be treated as if it were credible: e.g., whether the current justification of the intuition prevents or invites bad consequences, vices, or duty-breaking. For instance, it might depend on the probability of the militant abolitionist's disposition to engage and endorse evil himself as compared to the divine command theorist. If the DCT believed murder was wrong because his God said so, but slavery is moral because God said so, we might feel similarly toward the justificatory basis of the two individuals' noncredible intuitions, and seek to extinguish the respective justificatory basis for their moral intuitions – self-interest and divine command, respectively – even though this would lead to extinguishing their proxy-credible moral intuitions (though they might be led back to these intuitions via other avenues; then again, they might not).

Consider proxy-credibility in relation to the subject of kin preference. Suppose a mother knows she has enough money either to buy a winter coat for her daughter or to make a donation to an overseas medical facility that inoculates vulnerable poor children against malaria. The mother feels more strongly obligated to provide her daughter a winter coat than to help distant strangers, so she feels morally justified in buying the winter coat (and might have felt it would be immoral to pass over her child in order to provide for distant strangers). The mother's moral intuition of kin preference, which we commonly take as self-evident, may be impugned in terms of credibility, given the biological explanation of kin selection.[253] We can then try to determine whether or not this kin preference is justified on auxiliary grounds. Presumably, MWRE will provide us with an ethical theory, or perhaps more than one theory, which is highly coherent and can provide reasons that justify kin preference in such cases.

In providing auxiliary justification, consider the morally relevant features of her child: both relational and nonrelational. The child is vulnerable, the product of the mother -- which presumably indicates some maternal responsibility, in need of provisions and nurturance, and the mother is person best positioned to provide and nurture. This prescription that she should provide a winter coat to her child is presumably derivable from moral principles, which have been constructed and/or corroborated by credible considered judgments. For instance, if a set of CJs support a set of principles similar to virtue ethics, then these principles would deliver certain moral prescriptions, such as "to be a good parent, and strive toward eudaimonia, a parent should show preference in care toward his/her offspring." – which we might view as a kind of corollary that is arrived at via moral principles.[254] Alternatively, rule utilitarianism, constitutive of a set of moral principles, could provide certain moral prescriptions that, if followed, would lead to maximized utility: presumably, utility would be maximized if mothers provided

[253] The evolutionary etiology suggests this initial judgment is error-disposed, due to genetic bias, and thereby has diminished credibility.

[254] To clarify: Kin preference, under this interpretation, is indeed a moral principle, but not one which is an abstraction from considered judgments as much as derivable from other substantiated moral principles: e.g., virtue ethics, feminist ethics, rule utilitarianism. By calling it a corollary, I do not wish to assert it is different from moral principles, but rather to emphasize that it is a moral principle of a certain subset: namely, those moral principles derivable from other moral principles.

favoritism to their children over those of strangers. The fact that this moral prescription, kin preference, is derivable as a moral principle from several different, and perhaps discrete, moral principles, further substantiates its credibility as a principle.[255]

Even though we might endorse a mother's having moral intuitions of kin preference toward her child, we might also constrain the extent of her kin preference. For example, if her kin preference prescribes buying her daughter a sports car rather than buying her stepdaughter a winter coat, or if she intuited she had more moral obligation to buy the latest fashionable clothes for her daughter on a weekly basis rather than to donate any money to Oxfam, we might conclude that the extension of her moral intuition in regard to kin preference outstrips its auxiliary justification.[256] This would extend to all parents who had this intuition, even if this intuition were cross-cultural and universal.

The discussion above, concerning a mother's care for her child, doesn't sound like a particularly heartwarming description of parental care. Our analysis represents a 3rd person perspective as MWRE methodologists, not necessarily as practical ethicists prescribing how individuals should approach morality in their lives. We needn't, and likely shouldn't, prescribe that the mother actually work through this reasoning herself. In fact, once we determine that the parent *should* have preference for her child, given auxiliary reasons, we might not object – and perhaps might endorse – that this mother maintain her intuitions of preferential treatment toward her child. Though her intuitions are not credible from the bottom-up as considered judgments, their forceful presence as moral intuitions might be an inexorable feature of being a parent. Even if these moral intuitions toward kin preference were extinguishable, it might still be a *good* thing that parents have these intuitions, and it might be a *bad* thing to extinguish these intuitions,

[255] In this way, a moral principle, if it is coherent not only with one's own WRE set but also the WRE sets of others, attains broadened credibility. If for instance, Kantian ethics and rule utilitarianism are both coherent with a moral principle, then this moral principle is further substantiated. The moral principle that the murder of innocent people is wrong, for example, is coherent with both the rule utilitarian's WRE set as well as the Kantian ethicist's WRE set, and thereby the moral principle gains broadened credibility.

[256] What exactly we do at this point is up for debate, and is a matter of practicality in regard to the moral costs and benefits of enforcing moral compliance. We can try to educate the mother in hopes of diminishing the intensity of her kin-preferential disposition, which would be "expanding the circle," as Singer phrases it. Or, if we may believe changing the mother's disposition is impossible – even if morally desirable – and thereby we might subsequently find other methods to ensure moral compliance, such as taxation; or, then again, we might take no steps to change her attitude or behavior.

especially if kin preference as a moral principle has systematized credibility and prescribes favoring one's kin over others.

Concerning a mother's moral intuition prescribing kin preference for her child (for instance, her buying her daughter an expensive and fashionable back-to-school dress rather than donating money to save 10 starving children), we need to consider how an evolutionary etiology might affect credibility. If the mother's moral intuition concerning kin preference arises, directly or indirectly, due to genetic relatedness, then without any other morally relevant substantiation, her kin preference intuition is not credible (at least on those grounds). Remember John and the experiment regarding individuals who are similar to him, phenotypically or genotypically. Such similarities were deemed not morally relevant in determining the distribution of resources or deserts. In this case of kin preference, it is more difficult to conceive of a relevant counterfactual, as a counterfactual would involve removing the ingrained genetic-ness of it, or the feeling of genetic-ness. So we would have to try to imagine that it were the case that the daughter were *not* genetically related to the mother, but the mother was more of a guardian of daughter and chose the daughter to somehow exist. In such a case, we might presume that the degree of the mother's moral intuition of kin preference would be lessened; that is, that the mother, bereft of the genetic-ness of it, would likely no longer have intuitions that extend so far as to providing for her daughter so disproportionately over the consideration of any others. The mother's intuitions of preference toward her daughter might be conceived as more akin to choosing to house a dependent, foreign exchange student. Preference toward this individual by the mother could be based on morally relevant features rather than genetically-generated normative sentiments: such as, voluntary bringing the dependent into existence, being in the best position to satisfy the individual's needs and interests, being best situated to help the individual flourish, etc. Despite lacking direct credibility as a considered judgment, kin preference intuitions should be afforded proxy-credibility via auxiliary justification to the proportionate extent of such justification. Determining the boundaries of that justification, we would reflectively place restraints that might reign in this preference: for instance, indicting the excessive kin preference of a mother who feels herself morally obligated to provide

daughter fashionable dresses to her daughter over providing for the dire needs of foreign children.

Our evolutionary etiology and its focus on kin preference might serve as a warning to parents: Not to feel as morally righteous or justified about kin preferences as they may immediately feel. After all, our intuitions regarding kin preference may be unduly influenced by genetic bias.

In summary, initial judgments that arise from the error-disposed condition of evolutionary bias are to be discarded from the set of considered judgments; however, we can attribute initial judgments "proxy-credibility" in that they can be treated as if they were credible, presuming they can be justified on auxiliary bases. This would make an initial judgment not credible as a considered judgment, but only proxy-credible via auxiliary justification.

Doing versus Allowing

In this section, I examine how a psychological etiology could diminish the credibility of a moral principle by providing a sufficient explanation for the presence of moral intuitions that might be taken to support it. Specifically, I consider a debate between psychologist Tamara Horowitz and philosopher Mark Van Roojen, which focuses on the doctrine of doing and allowing. Horowitz claims a psychological etiology, namely our disposition to employ prospect-theoretic valuation, undermines the credibility of the Doctrine of Doing and Allowing (DDA) as a *moral* doctrine. That is, the psychological explanation of the doing/allowing distinction may diminish the credibility of these intuitions as intuitions that map on *moral* features of the world.

The Doctrine of Doing and Allowing is a moral principle that, generally formulated, suggests that passively allowing harm is less immoral, all else being equal, than actively doing harm. For example, passively allowing a neighbor's child to drown in their front yard kiddy pool is immoral but it is still *less* immoral, all else being equal,

than actively drowning the child in the pool.[257] Similarly, passive euthanasia seems to be more morally acceptable than active euthanasia, all else being equal. Generally, actively causing harms to occur is thought to be morally worse than passively allowing the same harms to occur.

DDA is a moral principle that can become substantiated in one of two ways:

(1) DDA is constructed initially as an abstraction from a coherent[258] set of moral judgments

(2) DDA is a principle we already happen to subscribe to and, upon reflection, turns out to be supported by a coherent set of moral judgments

In either case, DDA – as a moral principle – is supported by the relevant moral intuitions.[259] In turn, DDA as a moral principle systematizes our relevant moral intuitions into a set. By systematizing this set under a singular principle, DDA confers consistency to this set of moral judgments. The anomalous intuitions – intuitions that are relevant to DDA but are not compatible with DDA – are bracketed and put aside, though they may remain credible considered judgments.[260]

One way to impugn DDA as a moral principle, generally, is to undermine the credibility of the moral intuitions that support it. The credibility of a certain set of moral intuitions is undermined if the following is the case:

---

[257] We would consider the bystander to be monstrous by omission, but less so than a "murderer" who actively drown the child in the pool.

[258] This coherence might be a matter of degree, as there often will be relevant intuitions that are not compatible with the principle.

[259] "Relevant" here means that the intuition refers to the same cases that DDA covers: that is, the intuition either confirms DDA as a principle or discredits it as a principle.

[260] Though we may think their lack of coherence indicative of some unexposed error condition that interfered with the generation of the intuition. Likewise, in science an anomaly is usually presumed to be due to some unexposed error condition. This presumption increases as the critical mass of the coherent set increases. Nevertheless, anomalies which are obviously not erroneous data can overturn a principle that is highly corroborated by a critical mass of data: for instance, consider the proposal that there was a moon orbiting Jupiter coined "Vulcan." This postulation was asserted in order to correct for a small hiccup in the prediction of Newtonian mechanics. It is astronomically obvious nowadays that there is no Vulcan moon. The anomaly, which astronomers originally tried to explain away via Vulcan conjecture, is a very fixed point of anomalous data – data which provided one foothold for the usurpation of Newtonian Mechanics by Einsteinian Mechanics.

(1) The determining etiology of the moral intuitions is error-disposed, *and* there are no other etiologies or MRFs evident that substantiate the credibility of the intuitions.[261]

In this case involving DDA, the error-disposed condition is that the determining etiology that generates the moral intuitions is completely unrelated to moral features of the object under consideration. Specifically, arbitrary "framing effects" significantly determine our moral intuitions regarding doing and allowing, which in turn diminish the credibility of these intuitions as considered judgments supporting DDA.

In illustration of the argument against DDA, we can consider Tamara Horowitz's 1998 paper, in which she asserts that a psychological etiology of the moral intuitions presumed to establish DDA as a moral principle, diminishes the credibility of those intuitions and, subsequently, diminishes the credibility of the DDA principle. Horowitz's discussion argues that prospect-theoretic reasoning (PTR) is a determining explanation of these moral intuitions. Prospect-theoretic reasoning states that "people tend to evaluate outcomes from some sort of neutral baseline and that positive deviations from that baseline are regarded as less significant than negative deviations" (pp. 847-848). Such reasoning seems unrelated to the moral features of the world. She concludes from this, that the aforementioned intuitions lose their credibility.

In illustration of prospect-theoretic reasoning, consider a notable experiment, called the "Asian Disease experiment," regarding "framing effects" presented by Tversky and Kahneman (1981). In this experiment, they presented one set of subjects the following scenario:

> "Imagine that the U.S. is preparing for an outbreak of an unusual Asian disease which is expected to kill 600 people. Two alternative programs to fight the disease, A and B, have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If program A is adopted, 200 people will be saved. If program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved. Which program would you choose?" (p. 454)

[261] Such morally relevant etiologies and MRFs might be searched for, depending on the nature of the determining etiology. I presume that the burden of proof should be pushed upon the individual who claims the moral intuitions are credible if it is shown that a morally irrelevant explanation sufficiently explains the presence of those intuitions. I will discuss this further later in the section.

Tversky and Kahneman then presented the same scenario to a second group of subjects, except that instead of being presented with programs A and B, these subjects had to choose between programs C and D:

> "If program C is adopted, 400 people will die. If program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 will die" (p. 456).

According to Walter Sinnott-Armstrong, programs A and C are equivalent, as are programs B and D (2008, p. 7). Interestingly, while 72% of subjects who chose between A and B favored A, only 22% of the subjects who chose between C and D favored C. Generally, subjects indicated they were "risk-averse" when results were described in positive terms, such as lives saved, but "risk-seeking" when results were described in negative terms, such as "lives lost" or "deaths."

Horowitz describes this phenomenon, generally:

> "In deciding whether to kill the person or leave the person alone, one thinks of the person's being alive as the status quo and chooses this as the neutral outcome. Killing the person is regarded as a negative deviation…. But in deciding to save a person who would otherwise die, the person being dead is the status quo and is selected as the neutral outcome. So saving the person is a positive deviation…" (1998, p. 153).

Programs A and C, and programs B and D, are both identical in content; the only difference is the way in which the scenarios are described. Though two programs are identical in content, the moral intuitions of subjects regarding the best moral action to take differ based on descriptions: namely, how content-identical scenarios are *framed* in regard to the baseline or "status quo." The way that content-identical descriptions in such cases result in significantly disparate moral intuitions of the subjects should lead us to question that credibility of the resulting intuition in such cases.

In relation to prospect-theoretic reasoning, if we knew that (1) prospect-theoretic reasoning determines our intuitions that support DDA, (2) that prospect-theoretic reasoning is independent from the moral features of the actions relating to those intuitions, and (3) that there are no other determining etiologies, then we could reasonably ascribe the initial intuitions less credibility, regarding moral normativity. That is to say, if the apparent etiology for the intuitions that support DDA is an etiology that is blind to morally relevant features, and there is no evident justifying etiology

present that supports those intuitions, then the credibility of the DDA principle is diminished.

If PTR is an etiological explanation regarding human psychology, namely how we valuate losses and gains according to baselines (as explained by Horowitz), then this psychological etiology may provide reason to question the credibility of our moral intuitions in gain-loss cases. In particular, the fact we are psychologically inclined to valuate gain-loss cases according to prospect-theoretic reasoning – irrespective of whether the cases are moral in nature or not – may provide a reason to suspect that certain sets of our moral intuitions are at least partially based upon morally irrelevant features (such as framing effects) and so are not credible moral intuitions.

Tamara Horowitz claims that prospect-theoretic reasoning could be a sufficient explanation of our intuitions that support DDA. If PTR is the only sufficient explanation for the intuitions that support DDA, then the intuitions might be suspect and, in turn, DDA becomes suspect. PTR only undermines DDA if the etiological explanation it provides is irrelevant to the moral features of the objects (moral thought experiments or situations) under consideration. It seems that PTR would count as irrelevant. PTR is a general psychological tendency in human beings. It is a way of thinking that turns up in other fields of study, particularly nonmoral ones, such as economics. It could be hypothesized that PTR carries over into the moral realm.

In consideration of psychological etiologies, Mark van Roojen considers the role of nonmoral explanations relating to the ostensible normative distinction between doing and allowing. Responding critically to an article by Tamara Horowitz (1998), Roojen considers the normativity of DDA. He states,

> "…psychological explanations do not initially play a role in justifying normative conclusions. However, such explanations of our particular judgments can be relevant if they show that our reaching the initial judgments involved some sort of error. But showing that we are making such errors will require more than a psychological explanation of our beliefs, even [a psychological explanation] which does not rely on the moral principle we have used the particular judgments to support. A case will need to be made that the reasoning postulated by the psychological explanation is inherently fallacious" (1999, pp. 846-847).

Roojen correctly criticizes Horowitz's presumption that a sufficient psychological explanation of the intuitions that give rise to DDA would necessarily undermine the

credibility of those intuitions. The basic argument that Horowitz presents states that DDA is ultimately determined by prospect-theoretic reasoning (1998, pp. 847-848).[262] She suggests that it is prospect-theoretic reasoning that determines our intuitions which, in turn, support the principle of DDA; in this case, the features that determine these intuitions are not *moral* features in the thought experiments under consideration. Thus, the *psychological* etiology diminishes the credibility of our intuitions, supporting DDA.[263]

As Roojen explains, Horowitz is mistaken in asserting a psychological etiology that determines an intuition necessarily diminishes the credibility of that intuition; for one, the intuitions supporting DDA could be overdetermined.

Roojen, however, states the objection too strongly by claiming that Horowitz' prospect-theoretic reasoning explanation could not undermine the credibility of intuitions. The following could be the case regarding DDA. We could know that prospect-theoretic reasoning sufficiently determines our intuitions regarding the moral differential between doing versus allowing. We could *also* know that there were no other plausible determining explanations. That is to say, we would know that prospect-theoretic reasoning was the *only* sufficient explanation for such intuitions. Since prospect-theoretic reasoning is an etiology unrelated to morality, we could conclude that the credibility of the respective moral intuitions would be undermined.

To clarify, let us revisit an earlier thought experiment by Kosfeld, Heinrichs, Zak, Fischbacher, and Fehr (2005). You awake in the laboratory of the two evil scientists, yet this time you have no memory of your moral intuitions prior to your awakening. As before, the two scientists inform you that they implanted in your intuition cache intuitions

[262] Prospect-theoretic reasoning states, "people tend to evaluate outcomes from some sort of neutral baseline and that positive deviations from that baseline are regarded as less significant than negative deviations."

[263] In an effort to support the plausibility of this claim, we might briefly consider an example. Imagine that sociological and psychological research supports that human beings tend to think attractive people are more morally deserving than less attractive people. Imagine that psychology offers the highly corroborated explanation that we have a tendency to morally valuate attractive people as more morally deserving than unattractive people because it's in our self-interest. In this case, it would seem that this psychological explanation, which is sufficient in explaining the presence of our disparate moral evaluation, diminishes the credibility of those initial intuitions: that attractive people are more morally deserving than unattractive people. In fact, there is psychological and sociological research that suggests this.

supportive of some moral principle P. Knowing that one determining etiology (implanted by the hypnotist, for instance) for these supportive intuitions provides a sufficient explanation for the presence of those intuitions (which support moral principle P), it seems those intuitions are subsequently less credible as *morally* normative guideposts. After all, the etiology that determined those intuitions (e.g., the whim of the hypnotist) might have no relation to morally relevant features. The principle of parsimony would suggest assigning less credibility to the intuition: if one determining explanation is sufficient, and no other determining explanation is evident, we shouldn't multiply entities, or explanations, beyond necessity. If no other explanations are evident, after a diligent search, then the single remaining etiology can undermine the intuitions' credibility if that etiology is unrelated to the moral features of the world. Roojen (1999, p, 855) concedes "An intuition would be a *moral* intuition if it depended on a sensitivity to moral features of an example, whereas it would not be [a moral intuition] if it instead depended on a sensitivity to features that would make a difference in choosing between options even where no moral issue is involved in the choice." Prospect-theoretic reasoning, demonstrated in the Asian Disease experiment, suggests that certain classes of intuitions are sensitive to morally irrelevant features, such as framing effects; therefore, such intuitions lack credibility.

Biological Etiologies, Trust Intuitions and Credibility

The fact that a biological/evolutionary or social/cultural etiology determines an intuition does not necessarily undermine the credibility of that intuition. In support of this important point, consider recent studies investigating the effects of oxytocin upon human subjects. According to a study conducted by Kosfeld and associates (2005), the introduction of oxytocin via a nasal spray induces human subjects to trust others more than they would under normal circumstances. In a game of investors and trustees, where investors chose how much money to entrust to trustees, those investors sprayed with

oxytocin (as opposed to a nasal spray placebo) entrusted significantly more money than the control group.[264]

Oxytocin is a natural chemical present in human and nonhuman animals. In nonhuman animals, oxytocin is conducive to social attachments: male and female bonding after mating, mother-infant bonding after birth, and sexual receptivity. It's speculated that oxytocin may play similar roles in humans. Oxytocin is thought to also lower animals' natural reluctance to the proximity of others, enabling what is known as "approach behavior," which can prove beneficial to both animals.

Experimental evidence suggests that the secretion of oxytocin in human beings significantly increases the subject's intuitions that other parties are trustworthy. Typically, oxytocin is secreted in response to certain stimuli: the features of a newly-born offspring observed by its mother, the submissive body language of an animal who has entered another animal's territory, a series of body movements indicating sexual receptivity. These stimuli prove to be reliable markers for trustworthiness. In this way, the secretion of oxytocin does not automatically undermine the credibility of the intuition that an individual is trustworthy merely because the intuition is significantly determined by the secretion of oxytocin. The natural secretion of oxytocin results from reliable markers of trustworthiness. In this way, we should accept X's intuition that Y is trustworthy to be a credible intuition insofar as the secretion of oxytocin was in response to these markers. The fact that our intuitions are biochemically influenced does not necessarily affect their credibility.

The credibility of X's trust intuition would be diminished if it arose under error-disposed conditions. One of these conditions would be if the secretion of oxytocin was not in response to reliable markers, but was in response to experimental or experiential features that were either unrelated to trustworthiness or purposefully deceptive. In Fehr's experiment, the secretion of oxytocin into the subject's nostrils represents an experimental feature that is unrelated to reliable markers of trustworthiness.[265] Whether

[264] One-half of the oxytocin entrusted all of their money to the trustees, as opposed to only one-fifth of the control group; of the remaining subjects, most of the oxytocin subjects entrusted a majority of their money, whereas only one-third invested the majority of their money.

[265] One could point to externality, artificiality, or other characteristics, which would require

or not subject Y is actually trustworthy, the oxytocin is "secreted" regardless. Neuroscientist Antonio Damasio of the University of Iowa College of Medicine in Iowa City remarks that we shouldn't be overly concerned with the actual abuse of oxytocin in everyday life, such as in politics or sales.[266] He suggests we need be more concerned about advertising and marketing which exploits our natural markers, thereby stimulating the secretion of hormones that unduly influence our intuitions, such as trust.[267] It's a conventional cliché, for instance, that politicians publicly kiss babies in order to influence onlookers to presume the politician is caring and trustworthy. Of course, these photo-ops are canned, staged, and typically insincere – their purpose being to manipulate voters into trusting the politician. This kind of etiology undermines the credibility of the subsequent trust intuition.[268]

Presumably, the instincts that nonhuman animals have concerning whether or not another animal is trustworthy is largely an evolved capacity. Those animals that had a reliable capacity to discern trustworthy from untrustworthy animals would have a selection advantage over those that did not, or that possessed such a capacity to a lesser extent.

Evolved capacities can be taken advantage of, by exploiting markers in order to manipulate subjects, as illustrated in the aforementioned example of political advertising to engender intuitions of trust in voters toward a particular political candidate. Evolved capacities can be reliable indicators at one point in evolutionary history, but become

---

specification. I will later examine why the experimental etiology is not credible whereas the natural etiology is credible. For now, I believe it's sufficient to point out that empirical tests would at least establish the reliability of intuitions resulting from the latter.

[266] He imagines the example of a politician spraying a crowd with oxytocin before she gives her speech. Also, we might imagine a used car salesman spraying oxytocin on his customers, in order to sell his vehicular lemons at a higher price.

[267] Other possibilities include advertising which exploits anger, fear, insecurities, and so forth. Political ads oftentimes employ demagoguery to convince people to vote for one candidate over another: for example, the infamous political commercial during the Johnson-Goldwater presidential election, where a little girl was depicted plucking the petals off of a daisy just before a nuclear bomb exploded nearby. The implied message was that if Lyndon Johnson was elected, a nuclear apocalypse would be imminent.

[268] Empirical studies show whether certain markers are reliable indicators of trustworthiness. Interrogators, such as police detectives, depend on markers – such as eye movement, facial expression, body language – to decipher whether or not a suspect is telling the truth, lying, concealing information, and so forth. Colloquially, an untrustworthy person might be described as "shifty-eyed," though this is meant figuratively rather than literally.

outmoded or "vestigial," where the capacity was once truth-tracking but is now error-disposed.

Our intuitions of trust regarding individuals who exhibit the appropriate trust markers extend beyond their mere dispositions, but into their character as well. Our trust intuitions concern not just how they will likely act, but is a moral valuation concerning if they are a morally good person, deserving of good treatment. If the markers are artificially exploited, the oxytocin is secreted, and the trust intuitions generated, then we have reason to reduce the credibility in these resultant trust intuitions. We have reason to question our moral evaluation of the person – not only as trustworthy and reliable, but as a good person, worthy of our moral affection.[269]

The content of our intuitions isn't limited to predicting the consequences of our association with others (i.e., whether or not they will reciprocate or take advantage of us), but also regards moral valuation of their moral character. However, it seems safe to say that the reason our trust intuitions have arisen from such markers, insofar as these markers are evolutionary (and thereby presumably somewhat cross-cultural), is strictly due from the consequences of relying on these markers: namely, that such reliance has afforded a selective advantage.[270] Even though trust intuitions presumably evolved due to the consequences (of selective advantage) – ultimately in terms of genetic self-interest – the trust intuitions are still credible in that they are based on morally relevant features, such as proximate consequences (e.g., pleasure/suffering, social wellness, and so forth). Again, while the intuition that stranger S is trustworthy is ultimately based upon evolutionary selective advantage, the intuition itself – which is more robust than just an assessment of stranger S's disposition towards action – is credible as long as it is based upon morally relevant features, via a proximate causal justification.

Retribution

[269] In fact, we might very well take the etiology of the exploitation of trust-markers as evidence that not only undermines the trust intuition, but causes us to take the opposite point of view: that they are very untrustworthy.

[270] As opposed to other markers, an emotional/trust indifference to such markers that precludes the selective advantage provided by cooperation in trustworthy individuals).

Another example concerning evolutionarily inculcated intuitions regards retribution. Evolutionary psychologists speculate that retributive sentiments found in human beings originally arose because they offered a selective advantage to the individual's gene propagation. If an individual did not punish a "cheater" (or someone who was unfair) this cheater would be more successful than the "sucker" – the individual being taken advantage of. Allowing a cheater to remain unpunished in one's small kin-based community is detrimental to all of those individuals involved (except for the cheater). If X doesn't sacrifice some resources to punish Y after Y cheats, then Y will continue to cheat (getting his back scratched, while not scratching back, stealing, not following conventions/rules, etc.), which will detrimentally affect the selection advantage not only of X but X's genetic relatives as well as non-genetic associates who reciprocally assist X. In summary, those individuals who sacrifice some degree of valuable resources in order to punish cheaters would tend to propagate more genetic material than those individuals who would not make such sacrifices. Animals would have a selective advantage if they had instinctual, emotional, even moral intuitions disposed toward punishing cheaters.

In a recent trend of psychological experiments, game theorists studied the retribution reactions of human beings.[271] In one popular type of game, called Public Goods games[272], individuals punish noncooperators, even when it is against that individual's self-interest.[273] Individuals will spend or forego resources in order to punish subjects who act unfairly – even in scenarios where the individual is not directly participating (and has nothing to gain or lose from the possible outcomes).

---

[271] See Issac, R., et al. (1994); Janssen, M., et al. (2003); Andreoni, J., et al. (2003). These retribution intuitions could be both biological and sociologically influenced or determined. I am not meaning to suggest there is some immutable core of human nature.

[272] One of the simplest versions of this game is where four players form a group. The experimenter gives each member $20, and each must decide how much of that allotment to contribute to a common pool. Once all four contributions are in, the experimenter doubles the amount of money in the common pool, splits it into equal quarters and gives that quarter to each of the four players. It is advantageous for each player to be a "free-rider" and donate no money, as for each dollar they contribute, they only receive only 50 cents in return. A slightly altered version of the game allows punishment of free-riders at the end of the round, but at 30% cost to the player who decides to punish. Players will eagerly punish free-riders, even if they know they will not encounter this player in any subsequent rounds.

[273] In addition, it seems to be against their genetic self-interest, as the punishment benefits no relatives or associates that will benefit the individual, even via genetic material propagation, in the future.

In another game type called "the ultimatum game," two subjects anonymously interact in a singular instance. The first subject is granted an amount of money, contingent upon acceptance by a second subject with whom the first subject is to share. The first subject must decide how to split the money: e.g., 50/50; 70/30; 90/10. The second subject must either accept the percentage split, or must reject it, in which case neither subject receives any money. Surprisingly, second subjects tend to reject any offers where the first subject shares less than 20 percent, since receiving any money benefits them more than receiving none.[274]

In both types of games, investors do not choose the rational self-interested option. Rational choice theory asserts that a purely rational individual will choose the optimal self-interested option. The self-interested option is that option that maximizes the benefit to the subject.[275]

Why do individuals punish cheaters and misers, even when it is against the material interest of themselves and even, at times, their associates? fMRI studies of subjects' brains during such game studies reveal that punishing is pleasurable. fMRI observations of subjects who punish cheaters and misers show that punishing releases dopamine in regions of the brain of the punisher: that is, it provides pleasure to punish. Evolutionary psychologists speculate that the human brain is wired in such a way where punishment of noncooperators gives pleasure, as this pleasure motivates individuals to take actions that tend to benefit one's genetic self-interest. This evolutionary explanation is similar to the oxytocin example, where trust is engendered toward another – via the hormonal secretion of oxytocin respondent to certain stimuli – as it tends to be in one's genetic self-interest.

In this section, I will argue that many retributive intuitions are not credible as considered judgments, though they may be proxy-credible via principles or other

[274] The point of acceptability to the second subject does vary somewhat among cultures, typically ranging between a 50/50 split as a maximum and 80/20 as a minimum.

[275] It could of course be that individuals are not self-interested, and punish not just to secure the interests of themselves or their associates, but due to the morally relevant feature of the harm that befalls strangers who are the victims of cheating, or because of the unfair and unequal distribution created by the cheater's actions. Such reasons represented morally relevant features, yet in order for a retributive intuition to be credible it must be actually based upon these MRFs.

considered judgments. The argument presented in this section against the credibility of retributive intuitions can be represented as the following:

1. Scientific and social scientific theories suggest that genetic self-interest sufficiently explains the presence of retribution intuitions in human beings

2. If a nonmoral etiology sufficiently explains the presence of a normative valuation, then the introduction of an additional explanatory entity, namely a moral explanation, is less credible[276] – unless there are morally relevant features evident upon which this additional moral explanation can be based (in which case the intuition would be overdetermined).[277]

3. In many cases where retribution intuitions arise, there are no apparent morally relevant features upon which a moral explanation could be based that would justify retribution intuitions.[278]

4. Therefore, in those cases, the retribution intuitions are less credible as considered moral judgments.

Stated another way: If evolutionary selection advantage sufficiently explains the presence of retribution intuitions in a nonmoral way, then the explanation which posits the additional entity of moral reasons for the presence of retributive intuitions should be substantiated by the presence of morally relevant features, upon which this additional explanation could be based. Otherwise, bereft of a substantiated moral explanation, the credibility of the retributive intuition, as a moral intuition, is diminished.

A retributive intuition can be credible as a considered judgment if it is based upon any morally relevant features. For example, a school teacher might have the retributive intuition that a disruptive student should be punished. This moral intuition might be caused by the retributive sentiment that the student deserves punishment, yet at the same time be overdetermined in also being caused by the teacher's recognition that punishment

[276] As we saw in the case of framing effects and prospect-theoretic reasoning.

[277] The prior case of trust intuitions represented a case of overdetermination: Though a nonmoral etiology sufficiently explained the presence of trust intuitions, there were also other causes – namely certain proximate causes – which were morally relevant and thereby provided a sufficient moral explanation.

[278] Since there is a sufficient explanation of retribution intuitions, and no other overdetermining explanation is uncovered after some investigation, the burden of proof is then shifted upon the retributivists to provide reason to believe an additional explanation is necessary: namely, that there are moral features of the world upon which retribution intuitions predicate.

ultimately benefits the student and his peers, as it facilitates learning by deterring future disruption. If the teacher's retributive intuition was based merely on the sentiment that the disruptive student deserved punishment (say a firm wrap on the knuckles by a ruler), and was not based on any considerations of benefiting the student or his peers or anyone, then the teacher's retributive intuition would seem to be bereft of any substantiating MRF, and thereby would lack credibility as a considered judgment.

A retributive intuition – such as that cheater C ought to be punished – lacks direct credibility as a considered judgment, unless an MRF can be found. Even if an MRF is found lacking, a retribution intuition can find proxy-credibility via auxiliary justification. This argument is very similar to the kin preference argument, where kin preference intuitions might not be directly credible, yet still be proxy-credible via content-identical considered judgments or derivable via moral principles. In the case of retributive intuitions, if retributive intuitions are content-identical to considered judgments based upon morally relevant features or are deliverable via moral principles, then these retributive intuitions are proxy-credibility via this auxiliary justification.

Certainly retributive intuitions, like kin preference intuitions, are pervasive among human cultures and embody strong sentiments that enter human moralizing. Pervasiveness and sentimental strength do not substantiate moral intuitions, however; moral intuitions, to be credible as considered judgments, must be based upon morally relevant features. In many cases, retributive intuitions do not seem to be based upon morally relevant features, and so cannot be substantiated as considered judgments, though they can perhaps be substantiated via auxiliary justification.

One example of a retribution intuition concerns capital punishment of a criminal. Consider a simple example, similar to one Kant presents: Kyle murders a citizen of his island community.[279] It turns out that all the members of the community had already unanimously decided to disband to the four corners of the earth, never to return. Instead

[279] Kant (2002) presents a similar example where he states: "Even if a civil society resolved to dissolve itself with the consent of all its members – as might be supposed in the case of a people inhabiting an island resolving to separate and scatter throughout the whole world – the last murderer ought to be executed before the resolution was carried out. This ought to be done in order that every one may realize the desert of his deeds, and that bloodguiltiness may not remain on the people; for otherwise they will all be regarded as participants in the murder as a public violation of justice" (in Rachels, p. 137).

of punishing Kyle, they merely plan to leave Kyle on the remote tropical island by himself, where they know he will contently live for the rest of his solitary life.

A morally upright community member might have the moral intuition that Kyle should be severely punished – even possibly killed – irrespective of the fact the punishment will cause no benefit (presuming Kyle is incorrigible) and will only cause harm (to Kyle – and arguably to the punisher or executioner). Evolutionary psychology would provide the explanation that community members would feel sentimentally motivated to punish Kyle: namely, because evolutionary forces selected for an emotional system that motivates punishing perpetrators of destructive behavior, given the deterrent effect of punishment, which increases one's genetic selective advantage. The moral intuition that Kyle should be punished would remain despite the ostensible lack of any morally relevant features in this particular case (including that one will be harmed and no one will benefit).

Punishing Kyle, in this case, can nevertheless be justified from a top-down approach: such as, as a corollary derived from principles. For instance, a Kantian might reason that Kyle should be killed because we are thereby respecting his rationality by carrying out his maxim in reflexive relation to him. This might serve as a sufficient auxiliary justification of the normative content of the retributive intuition: in our case, that Kyle should be punished. Presumably, this is not a reason why we have the retributive intuition in the first place, however; therefore, this justification is auxiliary, which results in proxy-credibility of the retributive intuition.

In the original example Kant considers, he asserts that morality commands capital punishment: community members need to kill the island murderer to rid themselves of "bloodguiltiness," or else they themselves will be complicit in the original murder in a certain way. Kant provided a compelling auxiliary justification for capital punishment, deriving justification for our retributive judgment via a moral principle. Many proponents of capital punishment, however, appear to base their retributive moral conviction – that severely offending criminals should be executed – largely upon their intuition that severe criminals "deserve" severe punishment. The state should inflict harm upon a violator because he deserves it. It is an intrinsically just act, which needs to be carried out even if

the practice results in bad consequences and no benefits to anyone.[280] The subject of these intuitions might reason thusly: the violator's act was immoral, the violator himself is vicious, and he deserves punishment, even if this punishment has no presumed or expected positive consequences, and even has some acceptable degree of negative consequences.

As previously mentioned in regard to Public Goods games, evolutionary psychologists provide a compelling etiology as to why we feel strong moral condemnation against "cheaters": it has been evolutionarily advantageous to an individual's genetic self-interest to have such attitudes. In the case of two kinds of individuals, P and Q, if P incrementally has a negative disposition toward cheaters whereas Q does not, P will punish cheaters which, for the above reasons, will increase P's fitness, whereas Q will not punish cheaters, which will decrease Q's fitness. Thereby, Q will be at a selective disadvantage to P.

According to a model presented by evolutionary economist Herbert Gintis of the University of Massachusetts, "social groups with an above-average share of punishers are better able to survive events such as wars, pestilence and famines that threaten the whole group with extinction or dispersal," the latter of which lowers selection advantage.[281] The members of such social groups are, of course, not aware of or directly motivated by the evolutionary advantages of punishment; they are only aware that they find emotional satisfaction in punishing noncooperators.[282] Sigmund, Fehr, and Nowak (2002, p. 85) endorse an evolutionary model that suggests that: "our emotional apparatus has been shaped by millions of years of living in small groups, where it is hard to keep secrets."

In the empirical studies of cooperation games, these same researchers observed that "a lot of players show great eagerness to punish defectors. Participants seem to

[280] For example, punishment could have the good consequence of deterrence, or it could have the bad consequence of encouraging an escalation in violence or a degradation of society's values.

[281] Explanation of this model is provided in "The Economics of Fair Play" by Sigmund, Fehr, and Nowak (2002). Recent research shows that social groups of pigtailed macaque monkeys elect certain members to "police" the group and keep order by mediating conflicts, and if these key monkeys are removed from the group, disorder occurs (Flack et al., 2006). Presumably this lowers fitness for group members.

[282] As previous stated, this emotional satisfaction is corroborated by fMRI observations of brain activity in subjects, where punishing cheaters and the like releases dopamine in regions of the brain of the punisher: that is, it provides pleasure to punish.

experience a primal pleasure in getting even with free riders. They seem more interested in obtaining personal revenge than in increasing their overall economic performance" (2002, p. 87).

If evolutionary psychology does in fact provide convincing evidence[283] that our psychological dispositions toward cheating, which seem morally normative in nature, evolutionarily arose due to the selection advantage it afforded, then such an etiology could diminish the credibility of retribution intuitions in certain circumstances. Similar to the case of the doctrine of doing and allowing (DDA), the credibility of an intuition is diminished if both of the following criteria obtain:

(a) The presence of the intuition can be sufficiently explained via a nonmoral etiology that explains the *why* the intuition is there in the first place[284]: in this case, retribution intuitions can be sufficiently explained by genetic self-interest.[285]

(b) There are no apparent determining MRFs or morally relevant reasons that support the intuition. In this case, a retribution intuition lacks credibility as a moral judgment if there are no apparent MRFs upon which to base the judgment.

It's important to clarify that the argument I am presenting here disputes the credibility of certain retributive intuitions that lack MRF substantiation. This account I am providing, if successful, would undermine the position of one type of retributivist, who solely relies upon the credibility of moral intuitions in justification of their position. I believe the forceful and vivacious nature of retribution intuitions, rather than a further contemplative moral reasoning, drive many people to favor positions that promote individual acts of retributive violence as well as institutional perpetrations of violence, such as capital punishment. Nevertheless, I am open to the suggestion that acts and

[283] There is such evidence suggested in the proto-moral behavior of certain primates.

[284] The "why" here is meant in the ultimate, purposive sense. For instance, we could ask *why* we have a trust intuition in a certain naturally occurring case. One answer would be that we have this intuition because the secretion of oxytocin in the brain due to certain stimulus. However, this doesn't answer the question regarding *why* the trust intuition is present in the first place (just as it's not an appropriate answer to the question "Why are we here?" by tracing the inquiring person's ancestry back to the Mayflower).

[285] As pointed out by Robert Richards (1986) in regard to evolutionary explanations and moral explanations, the principle of parsimony suggests that if a sufficient explanation is known for a phenomenon, it is a less credible explanation to posit an additional entity, which is unnecessary to explain the presence of the phenomenon. If there are no apparent MRFs upon which an intuition can be predicated, then the credibility that it's apprehending some moral reality is diminished (for example, the intuition that incest is immoral when MRFs are absent)

policies of retribution, such as capital punishment, though perhaps not justifiable from the bottom-up via credible intuitions, are perhaps justifiable via moral principles.

Vestigial Intuitions

In this chapter we've examined several kinds of moral intuitions and their evolutionary etiologies, which bear upon their respective credibility. In order for a moral intuition to have credibility, it must be based upon a morally relevant feature. Through the development of human history, and the changes of civilizations, morally relevant features may have developed and changed as well. In this section, I will examine cases where MRFs have possibly changed. Consider two examples, previously considered: (1) Trust intuitions are credible when in ancestral environment or ordinary circumstances, but not in environments where there are stimuli created to exploit trust intuitions artificially (e.g., political advertisements for a candidate). (2) Retribution intuitions are credible in a tribal cultures, or evolutionary environments, where RI motivates maintenance of social functioning, for instance; however, retribution intuitions are not credible in certain modern-day contexts, such as capital punishment in contemporary America, where there are no apparent MRFs upon which to predicate these particular retribution intuitions (though these retribution intuitions may be allocated proxy-credibility).

In human evolutionary history, retribution intuitions supervened upon morally relevant features[286]: namely, RI arose out of the deleterious social consequences that resulted from not punishing cheaters. The etiology of retribution intuitions, then, is ultimately genetic, though proximately social in nature. In this way, we could say that RI supervened upon consequences: namely, that negative consequences to one's self-preservation and prosperity as well as to one's kin and reciprocal altruists are a morally relevant feature upon which RI can be predicated. Even if the ultimate reason for the presence of RI is genetic self-interest, the proximate reason, if it provides an MRF, will

[286] This is not to suggest that the reason the intuitions were present was causally *because* of these morally relevant features. However, we can recognize them as morally relevant features after that fact.

suffice in justification. However, if there is no MRF present upon which RI can be predicated, RI is no longer credible: the proximate causes are no longer are present, which provided MRFs.

Retribution intuitions may supervene upon morally relevant features: namely, morally relevant consequences. However, in significantly different circumstances, these underlying morally relevant consequences might no longer be present. There is good reason to believe that retribution intuitions, while once credible, now lack credibility in certain conditions, as there are no longer those MRFs upon which to predicate RI in those circumstances.[287] For this reason, we should consider RI, at least in these situations, to be a vestige of our evolutionary past.

To elucidate the argument relating to retribution intuitions via comparison, let us revisit the subject of trust intuitions. We should realize that evolutionary selection has provided us with typically reliable intuitions regarding what markers to take as indicative of trust. For instance, if a stranger smiles and meets you eye to eye, this is typically a reliable sign of trustworthiness. However, in particular circumstances, we might feel ourselves experiencing an intuition of trust, but realize we need to attribute less credibility to this intuition. For instance, if we are in a social circumstance where persons are trying to sell us products – and know they have been trained in exploiting natural trust markers[288] -- we would be wise to attribute less credibility to our trust intuitions.[289] While our intuitions based upon these trust markers usually remain reliable in our everyday lives, in certain modern-day contexts, we need to be prudent regarding the degree to which we should attribute credibility to our intuitions.

---

[287] As previously stated, retributive intuitions may be attributed proxy-credibility in that other MRFs may be appealed to, such as reasoning based upon Kantian ethics. These reasons, however, seem to be post hoc reasons that are not responsible for the presence of the intuition, and do not satisfy the counterfactual test, though these reasons do provide some indirect warrant to accept the intuition (at least partially).

[288] Such markers include prolonged eye contact, repeatedly using a person's first name in conversation, intimating personal information about one's self, gift-giving, etc. Pharmaceutical representatives, for instance, use all of these techniques to gain the trust and favor of medical professionals. Though prohibited from directly giving large gifts, they take doctors out to lunch, offer them small gifts, such as pens, and try to build a personal rapport.

[289] I speculate that this might be one reason many of us feel more comfortable bringing a friend when we're in such situations, such as shopping for a new car: not only to help us not buckle under pressure, but also to resist being lulled into a deluded sense of trust to the very friendly salesperson.

Similarly, we realize that evolutionary selection has provided us with typically reliable intuitions regarding behavior which will cause us and our associates harm.[290] Retributive intuitions are a set of intuitions that (at least) partially supervene upon what will cause such harm.[291]

One contemporary context in which retribution intuitions may arise is in response to another individual's criminal actions, where a citizen believes society (or its citizens therein) should inflict injury to the criminal because the criminal simply "deserves" it. In our contemporary, industrialized society, the social form of life tends to be significantly different from the evolutionary environment: 21st century society is governed by criminal and civil courts, law enforcement, and a penal system. Given our system of imprisonment, which both deters and rehabilitates, directly inflicting harm upon violators for transgressions seems unnecessary and unbeneficial. In fact, inflicting harm may be detrimental to all parties involved.[292]

In this way, we might believe that while retribution intuitions were useful to have in our evolutionary past, and had some credibility as they supervened upon MRFs, they should now be considered, in certain contexts, an outmoded vestige of evolutionary history. It might be the case that among tribal cultures in the present day, RIs are still the essential mechanism by which to ensure social stability and prosperity; in this context, they would remain credible as moral intuitions. Temporality is not at issue here, but correspondence between intuitions and environment.

In the modern, industrialized world, RI intuitions can be provided proxy-credibility via auxiliary justification: such as derivation as a moral principle via other moral principles, or to the extent that RI is content-identical to other considered

---

[290] A fact which I presume to be morally relevant, not only on the basis of pain/suffering, but on the basis of autonomy, liberty, and other morally compelling features.

[291] We can understand partial supervenience to be a case where there usually is an association. For example, bright, lustrous coloration in male peacocks – in natural environments – is a supervening property of reproductive potential. However, this is only a strong association, but isn't always the case: you could potentially have these features originating from accidental genetic or environment factors, thereby not connected with fitness. And you could have high reproductive potential without such coloration. However, such coloration is thought to be an emergent property of reproductive fitness: physical markers that are indicative of genetic fitness.

[292] For example, initiating/perpetuating an endless cycle of violence and vengeance; degrading the lives of the executioners; complicating international cooperation, and so forth.

judgments available that actually do predicate upon morally relevant reasons.[293] Many retributive intuitions, nonetheless, cannot supervene upon vestigial presumptions of deterrence, since it is significantly no longer the case. So when faced with the issue of capital punishment, to the extent that evolutionary psychology accounts for the strong presence of those intuitions, while there is an absence of other MRFs determining the intuitions, we should be less confident in the credibility of that intuition. That is, in the absence of any apparent MRFs to support the moral conviction that criminal X should be put to death, we should be less moved by mere force of our resonant intuitions to feel our position is justified, as the intuition is ostensibly unsubstantiated. Again, other accounts may be proffered in justification; indeed, the burden of proof is shifted onto retributivists to provide such support.

Our retribution intuitions are similar to our compelling intuitions regarding kin preference, and for the same ultimate reason (though different proximate reasons) of genetic self-interest. Genetic self-interest, however, is not a MRF. The credibility of our RIs is only substantiated if they are predicated upon MRFs. I have proposed that the individual and social detrimental consequences resulting from a disposition to permit cheating suffice as morally relevant features. The presence of such MRFs warrant the credibility of our retribution intuitions that morally condemn cheating; however, insofar as this MRF – or any other – is absent in the case of capital punishment, the supporting intuitions for capital punishment will lack credibility.

Credibility and Correspondence

In order for an intuition to be credible it cannot arise under error-disposed conditions and it must be predicated upon a morally relevant feature. A moral intuition that is originally credible, as it predicates upon morally relevant features, can later lose its credibility in cases where those MRFs no longer obtain. In relation to vestigial

[293] Such as philosophical arguments that rely upon a Kantian notion of respect for persons, which justify punishment as respecting a person's decision via appropriate and foreknown retributive response.

intuitions, let us consider a non-evolutionary hypothetical example where a moral intuition begins as credible, but later loses some measure of credibility:

Thomas and Marie are both in their twenties. They live next-door to each other. Thomas negligently backs up over Marie's cat. In addition, Thomas refuses to make any apologies or amends to Marie. Marie has the moral intuition that Thomas is extremely blameworthy and needs to make heartfelt amends, admitting he was wrong for what he did and providing some measure of appropriate compensation. Sixty years pass, and Tom (who, in his relaxed retirement, has shortened his name) and Marie still live next door to each other. Both are now in their twilight years. Tom seems vastly different from the young adult, Thomas, who accidentally ran over Marie's cat.[294] Nevertheless, Marie, steadfast and unforgiving, still generates, to the exact same degree, the moral intuition of Tom blameworthiness.

Isn't Marie's moral intuition of Tom's blameworthiness outmoded, at least to a degree, and thereby less credible? We might think that Tom, being contiguous in some way with a much younger Thomas, perhaps should apologize and make amends with Marie. However, if we judge him to be so obligated, it would be in a much different way. We might even liken Tom to a responsible parent apologizing to a wronged neighbor for his child's transgressions; we would hardly think elderly Tom to be blameworthy in the same way and to the same degree as young Thomas was blameworthy. If Tom came to Marie's door (presuming he suddenly remembered the incident he'd forgotten about so many years before), we would hardly agree with Marie's reaction -- one of seething resentment and moral condemnation of Tom (though that would seem to have been appropriate for the young Thomas many years before).

In this way, we can see that Marie's moral intuition in regard to elderly Tom is a vestigial intuition that has become outmoded. There no longer is an (as) appropriate object for her moral condemnation. Elderly Tom isn't nearly the same exact object to which Marie's moral condemnation corresponds (whether the condemnation presumes Tom's moral blameworthiness, disposition toward disutility, vicious character,

[294] This could even bolster this example by strapping Tom with psychological conditions where his personality has drastically changed, his old memory erased and a new one implanted, and so forth.

irrationality, corrupt personality, etc.). We might think it's silly, then, for unforgiving Marie to maintain her intuition that Tom is still as morally blameworthy now as he was 60 years ago, when he was just a young adult. Marie might maintain that her intuition was just as justified now as it was then; we, however, if we were cognizant of the back-story, would reasonably attribute diminished credibility or validation to her intuition.[295]

Another example we could consider is the moral prohibition on eating pork, by certain religions. Social (as opposed to biological) forces inculcated in children the intuition that the act of eating pork is intrinsically immoral. The reason that this intuition was initially inculcated was presumably because during that historical time-period, eating pork resulted in ill-health and oftentimes death. In that context, we might imagine that the intuition "It's immoral to eat pork" had at least some proxy-credibility: that is, while the intuition was indoctrinated by parental enforcement, rather than upon predication on anything in intrinsic to eating pork, there was an MRF present: namely, the *consequences* that resulted from eating pork (even if the negative consequences aren't the factor upon which the children's intuitions are predicated). Nowadays, however, given more healthy conditions, safer food preparation, and medical access, the intuition that asserts eating pork is intrinsically immoral no longer has even proxy-credibility.[296] Interestingly, the pork prohibition intuition was based *ultimately* on the MRF of bad consequences (as that's why the intuition arose in the first place), but *proximately* upon the morally irrelevant feature of parental inculcation.

Selection Advantage and Morally Relevant Features

The following is a speculative discussion exploring the connection between selection advantage and morally relevant features. Empirical substantiation would need

[295] Identity, not temporality, is the issue here. We might imagine, for instance, if young Thomas were cryogenically frozen after his transgression, only to be unfrozen 60 years later. In such a case, unforgiving Marie's intuition would seem justified, as the object of her intuition is quite the same object that negligently and apologetically ran over her cat.

[296] One could assert that the moral obligation is in reference to divine command or perhaps even moral fidelity to one's heritage. The example however regards the intuitions in, say, young adults who do not investigate or consider to a significant extent upon what basis this prohibition is justified.

to be provided to bear out the correlation I am proposing to make here. However, nothing too significant hangs on this correlation. My assertion, which seems at least plausible, is this: If we empirically examine purported evolutionarily adaptations in human beings, we will tend to find a close correlation that, in the evolutionary environment, if something is selectively disadvantageous, it is tends to cause pain/suffering/dissatisfaction/injury, and if something is selectively advantageous, it tends to contribute to pleasure/happiness/satisfaction/flourishing.[297]

It is not my intention to assert that moral norms are justified by natural selective advantage, and immorality illegitimated by selective disadvantage. I am merely pointing out a conjunctive tendency between the presence of selective advantage and the presence of morally relevant features. I am taking morally relevant features to be justified, as previously discussed, by a pragmatic stance or an overlapping consensus or how we must understand morality, rather than on the basis of whether or not human actions are natural or biologically normative.[298]

I am suggesting that morally relevant features roughly track evolutionary advantage, so if something is evolutionarily advantageous, then we are more likely than not to discover the presence of morally relevant features. A mother's love, for instance, is evolutionarily advantageous to both the mother and to the child; in fact, its presence is ultimately explained – at least partially – by the evolutionary advantage it affords. Nonetheless, I am not claiming that maternal love is not positively morally justified because it evolutionary advantageous or "natural." There is only an approximate tendency for evolutionary selection advantage to match up with MRFs: metaphorically put, where there's fire, there's usually smoke.

If it is the case that MRFs tend to emerge from evolutionary selection advantage, then if an intuition affords an evolutionary selection advantage, we can presume that it is

[297] Eating food, for instance, causes pleasure, whereas hunger causes pain. Bodily injury causes pain; sex causes pleasure; and so forth. There are, of course, evident counterexamples: Pregnancy and labor causes extensive pain and suffering though it results in further selection advantage via propagated genetic material.

[298] The justification of "morally relevant features" was discussed previously in chapter one: morally relevant features are those features recognizable by some ethical form of life, which is constrained in the ways presented by Philippa Foot (1958; 1959).

likely it also supervenes upon morally relevant features, under normal circumstances.[299] In the case of retribution intuitions, we can presume that retribution intuitions are present in humans because of evolutionary selection advantage it affords; additionally, we can presume that under normal conditions, we will also find morally relevant features to be present. Contrariwise, we cannot be so assured that, under *abnormal* conditions, we should expect MRFs to be present.

A Second Argument

This normal/abnormal distinction, in regard to intuition generation conditions, provides a second possible argument that can be presented in regard to vestigial intuitions, asserting that novel contexts tend to correlate with diminished intuition credibility. To illuminate the argument, let us revisit the oxytocin case. Oxytocin is secreted in the brain when the individual perceives certain markers of trust in other individuals (such as body language); as a result of oxytocin secretion, the individual is influenced to generate trust intuitions. These intuitions are trustworthy in an evolutionary environment in which they arose under selection pressure; however, these intuitions are no longer trustworthy (or credible) in a drastically changed environment, for instance, where the markers are artificially presented in order to exploit trust intuitions (such as political commercials depicting politicians kissing babies).

Similar to the oxytocin case, when our retribution intuition arises in the evolutionary or natural environment, we can presume it to be more credible than in novel circumstances. An intuition tends to be more reliable if you don't depart from the non-error-disposed context in which it was originally inculcated.[300] In proportion to the departure from the original context is the multiplication of error possibilities. Oxytocin, if it is secreted due to natural markers is probably bound to be more reliable in that original environment than some new environment which has unanticipated and abnormal

[299] Such as circumstances that would occur in an evolutionary environment, where the intuition arose due to selective advantage.

[300] This argument is similar to the argument Hare presents (1981, ch. 2): that moral intuitions are credible in those situations for which they were inculcated to handle, but are far less credible in novel or artificial contexts. I will discuss this argument further in the next chapter.

conditions. Essentiality my claim is that the simplest circumstances, without superfluous and additional entities that can interfere with fidelity, provide the most reliability.

Responding to Reductionism

Several scientific and philosophical thinkers assert that an evolutionary etiology that sufficiently explains the presence of moral intuitions thereby automatically undermines the normativity of those moral intuitions. Even if morality were entirely determined biologically[301], this would not undermine its normativity. Consider the moral intuition that killing one's brother is wrong. An evolutionary explanation of the presence of this intuition would cite that it's against an individual's genetic self-interest to kill his brother. Despite the fact it's against genetic self-interest for an individual to kill his brother, and presuming the counterfactual holds that if it were *not* against genetic self-interest, then that intuition would be absent, these factors needn't undermine the credibility of a moral intuition that killing one's brother is immoral. Even though the moral intuition arises ultimately because of genetic self-interest, as long as the moral intuition is predicated upon other more proximate morally relevant features or reasons – such as suffering, rational consistency, autonomy, etc. – the intuition will remain credible.

We can imagine a more extreme dosage of moral reductionism which asserted that all moral intuitions are just products of evolution (and social forces), and therefore have no morally normative force; in other words, these seemingly "moral" intuitions are merely biological imperatives disguised as morally normative imperatives. The moral reductionist could continue that if it weren't for genetic self-interest, one might have no moral qualms about killing his brother: for instance, if an individual's genes were better propagated through killing kin.[302] In such a case, we might imagine that individuals that

[301] "Determined" here is not referring to biological determinism, which asserts that biology strictly fixes how we behave.

[302] We might imagine this could be the case in a circumstance where there was a surplus of males and very few females available for gene propagation. In such a case, a male individual's brother would transfer part of that individual's genotype indirectly; nonetheless, direct reproduction would be significantly more successful in propagating one's genotype. If this circumstance were a fixed evolutionary circumstance,

evolved under such circumstances would have vastly different intuitions from the ones we have. E. O. Wilson (1975, pp. 198-199) speculates in a similar way about our moral valuation of human rights and freedom: "A rational ant – let us imagine for a moment that ants and other social insects had succeeded in evolving high intelligence – would find such an arrangement biologically unsound and the very concept of individual freedom intrinsically evil."

While this is true – vastly different evolutionary circumstances would result in vastly different moral intuitions – I don't believe it proves problematic. In analogy, if we evolved to be like bats, our subjective experience of the objective world would be entirely different, phenomenologically. Thomas Nagel (1974) argues that what it is like to perceive like a bat is inconceivable to us in the current positions in which we are situated: namely, our being situated as human beings with perceptive senses different from those of bats. This strange contemplation that bats perceive the world differently from human beings – for instance, perceiving material objects and obstacles via sound rather than sight – doesn't seem to impugn the fact that when I see a table before me, I'm confident in calling it rectangular, hard to the knocking, smooth to the touch, and so forth. It seems unreasonable to indict these perceptions as noncredible merely due to the evolutionary determination of my sense faculties.

Regarding our discussion, all that can be taken from such extreme examples is the fact we might have different moral intuitions, had evolution been different. Presumably, along with that vast difference in evolutionary history would be a vastly different "human" condition; for example, what brings us pleasure/pain, fulfillment, satisfaction of whatever our interests happen to be, would be different. Even if these things were identical to those we have now, this worry doesn't seem like an insurmountable problem. For instance, consider if a "human version 2.0," had evolved the moral intuition that killing one's brother was morally good – similar to Darwin's example of an intelligent hive of bees. The intuition morally commending fratricide, though evolved, would still be noncredible as it is based upon no MRFs, whereas the contrary prohibiting fratricide does

---

then males might evolve moral intuitions permitting killing kin – or even encouraging it as a moral imperative.

in fact predicate upon MRFs. I would consider this case to be similar to the intuition human beings do in fact seem to have, or have had historically, that killing or denying help to individuals in one's out-group is far more morally permissible than doing so to one's in-group.

Chapter 5: "Wide Reflective Adjustment: Defending Utilitarianism"

Introduction

The goal of this dissertation project extends beyond providing a method by which moral intuitions can be determined as credible and noncredible. The larger aim is to show how the moral methodology of wide reflective equilibrium (MWRE) can provide further resources in assessing ethical theories and adjudicating ethical debate. In illustration of this upshot, I will show that by employing the methodology of wide reflective equilibrium, a partial but significant defense can be provided to a version of utilitarianism.

The method of wide reflective equilibrium can reach coherence in three ways: vetting moral intuitions as credible or noncredible; adjusting moral principles and theories in reflection of considered moral judgments, and vice versa; and adjusting moral judgments and moral principles in coherence with background theories.[303] All of these adjustments work in tandem with one another, though I will consider them relatively separately in this chapter.

To show MWRE at work, this chapter will demonstrate all three coherence adjustments noted above. However, given that chapters three and four primarily focused on determining the credibility of intuitions, this last chapter will focus on theory adjustment. Specifically, the lion's share of this chapter's discussion will be dedicated to providing a more sophisticated version of utilitarianism, which I term "dispositional utilitarianism." I proffer this revised version of utilitarianism not as an *ad hoc* theory adjustment, but as a natural adjustment of a utilitarian moral theory given background

[303] Background theories may also be adjusted in MWRE, depending on its degree of substantiation. Oftentimes background theories enjoy independent justification, such as scientific grounding, and so are relatively resistant to adjustment in MWRE. Moral background theories, as discussed in chapter three, may be one subcategory of background theories that are more prone to adjustment in MWRE than most nonmoral background theories. Nonmoral theories, such as metaphysical background theories, are vulnerable to adjustment in MWRE: one possible example may be the problem of evil: moral intuitions and moral principles – that suffering exists, which is intrinsically morally bad,, and the moral principle that any good agent would not create/allow evil when she could easily prevent it. The critical mass of moral intuitions and moral principles could oust our nonmoral (metaphysical) background theory that a perfect deity exists.

theories relating to the human nature of moral agents with certain projects, commitments, and character dispositions.

Traditional attempts to discredit act utilitarianism tend to rely upon conflicts between considered judgments and the theory's alleged prescriptions. Interestingly, many of the troublesome moral intuitions dogging act utilitarianism are similar ones to those impugned in the last chapter: intuitions relating to doing versus allowing; kin preference; and retribution, among others.

One simple way act utilitarianism can evade such conflicts is by dissolving the contrary moral intuitions by impugning their credibility via filtration. Reasons for or against an ethical theory, after all, must be *good* reasons, and intuitions – insofar as they constitute reasons to accept or reject an ethical theory – must be credible intuitions.

However, an essential part of MWRE is that a moral theory, itself, can be revised in order to allow for greater coherence with moral judgments and background theories. For instance, many traditional objections to act utilitarianism theory appear to critique relatively unsophisticated versions of act utilitarianism, while not considering whether or not more sophisticated versions of utilitarianism can surmount such objections.

Intuitions versus Ethical Theories

Ethical theories are subject to intuitions: intuitions corroborate and test moral principles and theories. Kantian ethics, for instance, must reconcile itself with the counterintuitive example of the inquiring murder.[304] Many articles have attempted to meet this objection from a Kantian – or neo-Kantian – paradigm.[305] The salient point is that Kantian ethicists take intuitions seriously, and Kantian ethics is tested, objected to, and revised or reinterpreted in deference to these counterexamples that derive their strength from intuitions.[306] Some individuals assess Kantian ethics as terminally flawed, given such decisive counterexamples which, if accepted, cast Kantian ethics as a rigidly

[304] Neo-Kantian ethicists seem to acknowledge deference to intuitions; however, it's dubious that Kant would himself (1889). See "On a Supposed Right to Tell Lies from Benevolent Motives" (Kant, 1981).

[305] See Korsgaard, 1996b, pp. 325-349.

[306] These revisions oftentimes turn on how the categorical imperative should be formulated, or what constitutes the criteria for rationality.

absolutist ethical theory, blind to morally relevant features, such as the negative consequences of actions.[307] In this way, Kantian ethics, like utilitarianism, is subject to corroboration or disconfirmation via moral intuitions.[308]

One way to defend Kantian ethics against traditional objections such as the inquiring murderer would be to assess the strength of that intuition. If it turned out, for instance, that our intuitions lacked credibility in regard to that case, then that objection would no longer serve as a credible or compelling counterexample to the theory. Liberated of this counterexample, Kantian ethics as a theory would then be further supported in that a significant data point against the theory would have been eliminated. However, in the Inquiring Murderer case, it doesn't seem likely that our intuition would change, and traditionally Kantian ethics has been revised, or interpreted in such a way as to allow additional sensitivity to certain morally salient factors.[309]

In the case of utilitarianism, when counterintuitions are undermined, unfettering the theory from such objections reliant upon them, utilitarianism becomes subsequently strengthened. In addition, if the filtration process ends up diminishing the credibility of a significant proportion of intuitions which happen to snap at the heels of utilitarianism, leaving relatively untouched those intuitions which are neutral or affirming of utilitarianism, then we would have reason to think that not only is utilitarianism strengthened by being unfettered, but that this liberation serves as corroborative.[310]

---

[307] However, Kant did attribute moral responsibility for consequences to individuals who acted in violation of the categorical imperative: for instance, if you lied to the inquiring murderer who subsequently left and happened to run into your friend sneaking out the backdoor, murdering the friend, you would be responsible for those consequences (though not in the case of telling the truth to the murderer). Kant's discussion of this counterexample appears in his essay, "On a Supposed Right to Tell Lies from Benevolent Motives" (1981).

[308] Virtue ethics is likewise haunted by counterexamples: for instance, the possibility that the Mafioso can be good qua human beings and achieve eudaimonia. For discussion on this topic, see Rosalind Hursthouse's *Virtue Ethics* (1999).

[309] Korsgaard (1996) advocates a two-level theory to accommodate what she calls "ideal" and "non-ideal" cases, whereas Kant, she attributes, has a single-level view that is only able to accommodate ideal cases.

[310] This seems relatable to the logic of the Monty Hall problem (discussed earlier in Chapter 4). It would seem peculiar if, upon intuition vetting, the only intuitions that were discredited -- among a wide range of affirming, neutral, and counter -- were those hostile to a theory. It would be reasonable to assume that this suggests there is a *reason* why no intuitions discredited the theory: namely, because the theory maps well, or is truth-tracking in regard to some kind of moral objectivity.

In order to clarify how the method of wide reflective equilibrium is functioning here in vindicating utilitarianism against traditional objections via a fortified filtration of counterintuitions, it will prove useful to revisit the analogy between MWRE and scientific methodology. As set out in chapter 2, the method of wide reflective equilibrium shadows scientific methodology: moral intuitions parallel scientific data and ethical principles/theories parallel scientific hypotheses/laws. In conjunction with background theories, moral intuitions can test, corroborate or discredit moral principles/theories. I will briefly overview the moral/scientific parallel in the section below. After delineating this parallel, I will introduce three caveats: being cognizant of them will help clear up confusion concerning how we should regard our moral intuitions, and when we should, or should not, rely upon them. I will show that ostensible conflicts can be dissolved if the intuitions are revised given a greater understanding of these intuitions, which can result in coherence with utilitarianism.[311]

Wide Reflective Equilibrium in Science and Ethics

Scientific theories are, in part, proportionately as strong as their corroborating evidence and, contrarily and concurrently, as weak as the body of counterevidence. There are various ways that scientific theories can surmount counterevidence: (1) reconciling initially anomalous data within the theory, [312](2) revising the theory in such a way that it ends up compatible with the anomalies,[313] or (3) invalidating the counterevidence.[314] In a

[311] This revision is discrediting the interpretation of the intuition, not discrediting the intuition itself: just what the intuition is about.

[312] This involves reconceptualizing the evidence, which is, in a sense, a destruction of the previous anomalous data and the discovery of new evidence. For instance, if a discovery of what was initially thought to be a planet, which would violate Newtonian mechanics, turns out in fact to be a moon, which would align with Newtonian calculations, then the data, though the same in referent, becomes reconceptualized. In regard to ethical methodology, the dispositional set caveat, which will be discussed later in this chapter, asserts that the initial moral data needs to be reconceptualized, which oftentimes results in a reconciliation of the once anomalous data with an ethical theory.

[313] For instance, German astronomer Johannes Kepler's revision, in 1605, of the assumption that planetary orbits are circular, and instead postulating that planetary orbits are, rather, elliptical.

[314] For example, we could show that the experimental procedures were wrong, and thus the generated data is discredited or suspect: for example, if the sample has been contaminated or the instrumentation is miscalibrated. Or we might reclassify the data: for instance, if we discover the alleged counter-data are

similar way, in the moral methodology of reflective equilibrium there are various ways moral principles/theories can surmount counterevidence: (1) reconciling initially conflictive moral intuitions with the set of moral principles (2) revising the moral principles/theory in such a way that it ends up compatible with the conflictive intuition, or (3) invalidating the conflictive intuition. The focus of this dissertation project so far, has primarily been upon the last approach: invalidating the counterevidence, via credibility determinations of intuitions. In this chapter, in addition to intuition credibility determination, the focus will shift to investigate the first two ways in some depth, and show how this tripled adjustment can lead to coherence between moral judgments, moral principles, and background theories.

To attain coherence in science, meticulous attention is paid to how data is generated: for instance, ensuring that precise experimental/observational conditions are satisfied, thereby guaranteeing the validity of subsequently generated data. In ethical reasoning by Nielsen (1977) and Sencerz (1986), however, far less attention is paid to the conditions under which moral intuitions are generated. Similar to generated data in science, which is assessed in terms of credibility, intuitions serve as data points in ethics, and must also be assessed in terms of credibility.

Throughout this chapter, I will consider a few famous counterexamples to utilitarianism and examine the credibility of the intuitions upon which they rely. While the intuition credibility determination may not result in a complete invalidation of the intuitions involved in each example, it will at least diminish the degree to which the intuitions are credible. As a result, utilitarianism as a theory will be defended against such objections.

This defense of utilitarianism will be accomplished by indicting the credibility of the intuitions upon which the objections are based. To this end, I will focus specifically on three caveats of which we must be cognizant when assessing the strength of intuition-supported objections: (1) predication (2) dispositional sets (3) boundedness.

First, in regard to predication, a supporting intuition may lack credibility in that it fails to be based on morally relevant features. Second, the scope of the morally relevant

---

actually artifacts of the experimental set-up, rather than generated data.

features upon which the intuition predicates may be wider than is originally presumed: particularly, the intuition may be predicated upon the basis of dispositional sets in addition to agential actions. Third, the supporting intuition may be overextended, where we have reason to believe the intuition no longer reliably predicates upon MRFs. These considerations, caveats to the methodology developed so far, should help illuminate the practice – and malpractice – of leveraging intuitions against ethical principles and/or theories.

Dispositional Utilitarianism

To avoid confusion, I should clarify at the outset the version of utilitarianism I will be employing in this chapter: dispositional utilitarianism. This version is not married to any specific definition of utility, and in that sense can remain unfixed: utility can be preference-satisfaction, happiness, well-being, or qualitative/quantitative pleasure. I don't believe the arguments presented in this chapter hinge on committing one to any particular definition over another.

Concerning the maximization of utility, there are a few ways to formulate the prescription of act utilitarianism. One version of act utilitarianism prescribes that the agent performs that action that he reasonably expects will maximize utility: this characterizes the notion of "subjective rightness." Of course, it may turn out that what had been reasonably expected to maximize utility did not in fact turn out to maximize utility: while a subjectively right action, it is not "objectively right." Nevertheless, by this version, the agent should only be praiseworthy or blameworthy in relation whether his actions conformed in deference to what he reasonably expected to maximize utility, not what actually turns out to maximize utility. This is assuming that the agent is not negligent, lazy, or irresponsible in his utility calculations, but did his due diligence to determine the utility-maximizing course of action.

Act utilitarianism needn't be committed to prescribing that an agent be motivated in deference to utility maximization, however. In fact, that might be a terrible way of maximizing utility. Adams (1976) termed another version of act utilitarianism, called

"motive utilitarianism," which is more akin to my account of dispositional utilitarianism, prescribes that agents have the propensity to perform those actions that in fact maximize utility (pp. 467-469). They needn't be subjectively motivated by utility-maximization; in fact, it's likely that they will not be so motivated, but rather will be motivated by a variety of things: commitments, projects, feelings of care, moral principles, relationships, etc. These agents can then be assessed in relation to what extent they will in fact maximize utility over the course of their lives by acting from such motivations. This is similar to dispositional utilitarianism, where agents need not be motivated by utility-maximization, and are to be assessed in relation to what extent they would typically maximize utility over a lifetime under the circumstances in which their dispositional set has been formed.[315]

Utilitarianism should be naturally concerned about dispositional sets. Agential actions do not emerge from nowhere, but are manifestations of an agent's character: actions affect character, and character affects future actions. Act utilitarianism concerns actions and the associated probabilities of their consequences, and endorses that action that reasonably can be expected to maximize utility. In the same way, dispositional utilitarianism concerns the dispositional sets that a human agent can have, and endorses that set that reasonably can be expected to maximize utility.[316] I would argue that this version of utilitarianism should be considered uncontroversial, as I believe it to be a natural extension of act utilitarianism understood as concerned about the maximization of utility, even if agents are not act utilitarians themselves.[317] In fact, I would contend that,

[315] These circumstances are *typical* circumstances: for instance, it may turn out by fluke that a person's disposition ends up not maximizing utility though it typically would in fact maximize utility. This is related in a similar way to an animal being optimally "fit" to survive, but end up not surviving due to circumstance. Fitness could be described as relating to "typical" circumstances. I will explore this parallel in later sections.

[316] In illustration, imagine that a utilitarian were to choose to introduce into the world one of three agents: she would choose the agent with a dispositional set most likely to maximize utility: e.g., a 70/30 selfish/altruist over a 90/10 selfish/altruist. Likewise, act utilitarianism has an interest in prescribing that an agent undergo moral education or taking character-building actions increase her utility-maximizing dispositional set. This can be linked to Robert Adam's motive utilitarianism (1976) where motives ought to be inculcated that will maximize utility in the long-run. A form of utilitarianism that acknowledges dispositions would prescribe an agent take those actions that cultivate utility-maximizing motivations, and refrain from actions that erode these motivations.

[317] Of course there is more to dispositional utilitarianism than this, where it should be recognized as

given human nature, human agents should not be utilitarians themselves in any stringent sense, where they would perform actions based upon a utility calculus: dispositional or otherwise.[318]

This version of utilitarianism that I am presenting leads to a divorce of "right action," as traditionally understood by act utilitarianism, from what an agent ought, morally, to do. I will discuss this in detail in later sections. "Right action," in relation to act utilitarianism, is typically defined as that action that an agent performs that maximizes utility.[319] According to dispositional utilitarianism, oftentimes an agent should not perform the right action, defined in this way. Rather, an agent should possess and act from, or develop those dispositions which would lead to him performing those actions that, in sum, would maximize utility over that agent's lifetime.[320] I will revisit the concept of "right action" in detail in later sections, and will consider counterexamples to dispositional utilitarianism that relate to right action versus agential prescriptions. As part of that discussion, I will recharacterize right action in relation to dispositional utilitarianism.

This elucidation of dispositional utilitarianism illustrates the way a theory can be revised or refined, where the result is a reconciliation of the theory with conflicting data. This refined notion of dispositional utilitarianism, however, does not do all the work in overcoming the four traditional counterexamples. The other two types of reconciliation – invalidating counterevidence and reconceptualizing anomalies – come into play as well,

---

distinct from act utilitarianism, but its beginnings seem to be a natural extension of act utilitarianism.

[318] This version of act utilitarianism seems better equipped to meet the alienation objection: that the demands of utilitarianism require agents to sacrifice their ground projects and commitments at the altar of utility. It could still be the case that even dispositional alienation would require an agent to sacrifice ground projects in order to prevent suffering (for instance), even if this would alienate the agent from his projects/commitments. Dispositional utilitarianism has more resources to meet this objection, however; and it's not clear, at least to me, that alienation from one's projects is categorically negative when the demands of morality are extreme, and the overall benefits are overwhelming. I do not wish to engage the alienation objection here, however, as this defense of utilitarianism is only a partial one, and the objections against utilitarianism are numerous.

[319] An action is subjectively right if the agent reasonably expects the action will maximize utility; an action is objectively right if the action actually turns out to maximize utility.

[320] I do not say that an agent should possess or develop those dispositions which would lead to him performing the most right actions, in sum, as it could be that the way he maximizes utility over his lifetime is by always performing that action that results in the second highest net utility, no single action of which would technically count as the right action.

as will become evident in the following sections. I will illustrate the first of these types, the invalidation of counterevidence, in the discussion, just below, of our first objection to utilitarianism.

Jim and the Indians

In what might be the most famous counterexample to utilitarianism, Bernard Williams (1973) presents the reader with the following scenario: Jim must either kill one Indian to save 19 Indians, or refuse to kill the one, which will result in the death of all 20 Indians by Pedro's hands. All else being equal, utilitarianism would clearly prescribe killing the one to save the 19.

The primary thrust of the objection is the claim that killing the one Indian seems to lead to a violation of that agent's moral integrity. The utilitarian would see an agent's moral integrity as reducible to an additional cost, sufficiently represented in the utility calculus, which, it turns out, would be outweighed by the value of 19 deaths against just the one.

The objection suggests that by shooting the one villager, Jim would, figuratively speaking, have blood on his hands, whereas not in the situation of allowing the killing at the hands of Pedro. This is another way of saying that the person is morally responsible for killing, but is not responsible – or is not morally responsible to the same degree or in the same way – for omitting from saving. The crux of the counterexample seems to be that utilitarianism does not value moral integrity. Even if killing the one to save the 19 seems like the lesser of two evils, any adequate moral theory should value moral integrity itself, rather than treating it as merely reducible to a small part of the utility calculus.

The moral integrity objection seems to rest, at least partially, upon the doctrine of doing and allowing (DDA), discussed in chapter 4.[321] If we have reason to suspect that

[321] Chapter 4 examines the research presented by Tamara Horowitz regarding the doing/allowing distinction; her findings provide some empirical reasons to question the credibility of the intuitions supporting the DDA principle: namely, the supporting intuitions seem to be sufficiently explained by psychological factors, irrespective of whether the context is moral or amoral; in addition, no morally relevant features seem to emerge after diligent search upon which the DDA principle could be based. The strength of the objection is diminished to the proportionate extent that the credibility of the doing/allowing

DDA may be based upon irrelevant psychological factors rather than upon morally relevant features, we have reason to carry that suspicion to objections which are fortified by DDA. This critical treatment of the intuition that supports this example invokes the 1st caveat: predication. If the intuition seems to be based upon features that may not be morally relevant, then the credibility of this supporting intuition is diminished, and subsequently the strength of the objection itself seems to be diminished.

In regard to the case of Jim and the Indians, what happens when we bracket from consideration the moral distinction between doing versus allowing? What results is a recharacterization of the Jim's two options: (a) Jim decides to perform action A, which results in the death of one person. (b) Jim decides to perform action B, which results in the death of 19 persons. All things being equal in the situation, choice (a) seems morally superior to choice (b), as the numbers do matter. Furthermore, given the bracketed doing/allowing distinction, the complaint via moral integrity seems less compelling: after all, we have reason to doubt the credibility of this distinction which seems to undergird the moral integrity objection.

We might view the objection more generally, however, irrespective of whether or not his moral conviction is based on DDA. The objection might be viewed as a general criticism that utilitarianism fails to respect Jim's moral integrity, and the importance of this integrity to his ground projects. Utilitarianism would of course regard moral integrity merely in terms of utility calculus. I would argue that this general objection, however, is still dependent on the validity of a moral principle for much of its force. If Jim holds a moral principle X, and it is dear and fundamental to his identity and life's meaning, X needs to be somewhat compelling – at least if we are to understand moral integrity in terms of acting in accord with moral principles the agent takes to be of significant value. Consider, for example, if Jim's moral principle were that he would not kill highly attractive people, and a utilitarian scenario prescribes he kill the village beauty queen in order to save 19 average-looking villagers from Pedro. In such a case, utilitarianism would be alienating Jim from this fundamental moral principle (that he holds integral to

---

distinction as morally normative is brought into question. While DDA may not have been disproved, the burden of proof is shifted onto defenders of DDA to show the principle to be a credible one.

his identity and foundational to his life's meaning). That utilitarianism would require Jim to violate his moral principle in this case, even though the principle is foundational to his identity and meaning, would seem to recommend rather than discommend utilitarianism.[322] In such a case, it seems that if Jim is alienated from this deeply-held moral principle, then this alienation should not be viewed negatively, and certainly not as a strike against utilitarianism.

If this retooling of the Jim case is thought to be dismissible due to the introduction of a questionable, and frankly absurd, moral principle, this only reinforces the intended point of the above example: namely, whether or not we take the violation of moral integrity as a compelling objection seems to significantly depend upon the legitimacy of the moral principle from which the agent is being alienated. In the original Jim and the Indians case, if DDA is impugned as a moral principle, then the moral integrity objection – even more generally – will be proportionately impugned. If the objection is to be considered compelling for another reason, and not DDA, a supporting moral principle needs to be located that utilitarianism contravenes.

Another response to the original Jim and the Indians example can be offered by dispositional utilitarianism. Dispositional utilitarianism can recognize that moral integrity – grounded in adhering to moral principles with which one closely identifies – is often a necessary part of a productive dispositional set, even though moral integrity may preclude utility-maximization in occasional situations. Likewise, insofar as moral integrity is essential to ground-projects, and ground-projects are essential to one's own happiness as well as a productive dispositional set, utilitarianism would accept moral integrity given these ends. One might object that utilitarianism is still not respecting moral integrity intrinsically, but only instrumentally. This, however, seems a different complaint than that the original objection that utilitarianism does not respect moral

[322] One might take issue with this recasting of the Jim example, citing that this 2nd Jim would hold an immoral principle. I do not think this moral principle is immoral, however: it merely applies the DDA distinction in a limited subset of cases. Jim would kill the one Indian to save 19 Indians, to maximize the lives maintained; however, he would have an issue with killing the one beauty queen to save the 19 average-looking persons (even though, like the first Jim, he might still make the sacrifice, as utilitarianism requires). Jim believes in DDA obtaining only when it comes to pretty persons; otherwise, he does not recognize DDA as morally relevant.

integrity at all. Does dispositional utilitarianism require an individual to violate their moral integrity in every case where utility would be maximized? The answer is no, if utilitarianism is understood in this dispositional form.

An agent might still experience alienation of a certain kind. Again, my intent is only to provide a partial defense of utilitarianism, not a complete one. One type of alienation, for instance, may occur when an agent encounters a conflict where he knows the right action, defined (by act utilitarianism) as that action that he can reasonably expect to maximize utility in that instance, and yet he may realize he should not do the right action. For example, Jim may realize that shooting the one Indian is the action that will maximize utility. However, he may also realize that if he shoots that one villager, despite the maximization of utility in this instance, he will degrade his character and cultivate dispositions antithetical to future utility-maximization. In such a case, dispositional utilitarianism would prescribe that Jim not to do the right action: the action that will maximize utility in that instance. In fact, in that case, dispositional utilitarianism would prescribe that Jim do the *wrong* action, and refrain from shooting the one villager – as it would lead to a degraded disposition that would not maximize utility over his lifetime. In this case, Jim would knowingly do the wrong action, knowing in doing the right action would degrade his dispositional set which would preclude him from maximizing utility over his lifetime. Jim might even experience guilt about this – knowing he could have saved 19 lives – even though he also knows that he did the right thing in regard to maximizing utility in the long run. In this way, there is a dissociative quality that Jim experiences: Jim might suffer a kind of alienation from doing that action he recognizes as the right action in a particular instance, and that action he recognizes will maintain the most utility-optimal dispositions that are likely to maximize utility over his lifetime. Of course, part of Jim's refusal is based on Jim's conviction that killing an innocent person is wrong, and would be a violation of his moral integrity. Still, this is a case where Jim does the wrong action – by not maximizing utility in this instance – and yet acts as he ought to, as I will later argue: namely, Jim acts in a way that is to likely

maximize utility over his lifetime vis-à-vis dispositions.[323] I will argue later that the notion of "right action" should be redefined in light of dispositional utilitarianism.

Regarding its only instrumental respect of moral integrity, the objection to be compelling, would seem to need to be based, again, upon a moral principle that was substantiated via credible intuitions. In this way, it seems even the general objection concerning moral integrity requires some basis in a substantiated moral principle. Utilitarianism, if it is to be rightly criticized, needs to fail to respect moral integrity grounded in a moral principle that ought to be respected.

As argued above, Williams' objection seems to be based, again at least partially, upon the presumption that there is a morally significant difference between doing and allowing: killing is an action that is more morally significant, *ceteris paribus*, than allowing death to occur where one could prevent it. However, if there is no actual moral difference in the act itself between doing and allowing, or even *less* of a moral difference than we originally intuit, then moral integrity is less compelling as some irreducible moral property: rather, moral integrity seems to move closer to being considered "moral squeamishness." Of course, it is possible to justify the position that killing the one to save the 19 is the morally correct decision, and that the contrary is immoral, but such justification would need to proceed on the basis of a substantiated principle.[324] Generally, objections, such as the moral integrity objection, energized by the normative

[323] It might be the case that Jim realizes that he will degrade his moral character, leading him to produce less utility over his lifetime. Nevertheless, he may realize that saving 19 lives by his killing one will outweigh this loss in utility-production, even if he produces no future utility later in his life. I take up this difficulty later, showing that dispositional utilitarianism would still not endorse Jim being of the disposition where Jim could at least "easily" make this sacrifice. Dispositional utilitarianism might nevertheless endorse Jim to be of the disposition that he *ultimately* be willing to make this sacrifice, after sufficient internal struggle. After all, at some threshold of the number of villagers Jim would save by killing the one (who would die in any event at the general's hands if Jim refuses), it seems to be the morally correct action, to our commonsense morality, to kill the one villager. William's complaint seems to be at the "easiness" that act utilitarianism gives its answer, with insufficient heed paid to things such as moral integrity. Presumably, moral integrity is, in part, a disposition-related trait: not just what makes one's life meaningful, but what keeps one's outlook positive, and what makes an individual's identity remain "whole" rather than dissociated or fractured. Dispositional utilitarianism seems to be able to account for moral integrity in a way that simple act utilitarianism is not.

[324] For example, upon the principled basis of Kant's 2nd formulation of the categorical imperative: treating one villager as a mere means to the end of the other 19 villagers.

vivaciousness of intuitions related to doing/allowing distinction, will be impugned if the credibility of the supporting intuitions becomes diminished.

One way of understanding the doing/allowing distinction is in terms of internal states of the agent: that the moral difference lies in the agent doing the action. For instance, virtue ethicists might claim that the moral difference between doing and allowing is that a virtuous person would allow a bad consequence to happen before they would perform an action that would result in a comparable bad consequence. According to this paradigm, a person would be virtuous, as it would make them good qua human beings and contribute to their (possibly) flourishing as a human being. In a relatively similar way, I would argue that DDA is constructed from intuitions concerning the dispositional sets of the agent: an agent who is able to easily kill for the greater utility tends to have a worse dispositional set, in our ordinary circumstances, than an agent who is "squeamish," and therefore cannot as easily bring themselves to kill one to save 20. In this way, the doing/allowing distinction may be based on dispositional sets, a concept I will discuss at greater length in the next section. We should note an important point, however, that relates to the caveat of predication: the doing/allowing distinction may turn out to be a somewhat illusory one, where the distinction is not as morally robust as we've been evolved or socially inculcated to think. It may be possible to recognize the limits of DDA, and then alter our dispositional set in accordance. Yet, we also need to be aware that human beings have a limited degree of malleability and plasticity, in which case we need to recognize that there may be some inexorable, innate features of human psychology that must necessarily persist in order for a human being to flourish or have a good dispositional set.[325] I will examine this third caveat, dispositional sets, further in following section, in specific reference one of the most commonly-invoked thought-experiments in moral philosophy: the trolley case.

[325] It may turn out that an agent, in her current state, is too squeamish about doing, in contrast to allowing, and can still be an equally (if not more so) morally upright person (even capable of flourishing qua human beings) while being less squeamish. For instance, perhaps she is squeamish about killing one to save a million people, whereas an agent with an equal or greater dispositional set would not be so squeamish. Nevertheless, both agents would be squeamish regarding killing one person for three people. It's far easier to imagine that the agent is virtuous in the first case as than it is in the second.

The Trolley Example[326]

In this section, I examine the trolley example, previously considered in chapter 1. The two versions of the trolley example relevant here are the lever case and the footbridge case. In the lever case, the bystander must decide whether or not to pull a lever in order to divert the runaway trolley from a track that holds five innocent people, who would otherwise be killed on impact, onto a track that holds only one innocent person, who will then be killed instead of the five. In the footbridge case, also known as the "fat man version," the numbers are the same, but instead of pulling a lever at the side of the track, the bystander is on a footbridge and can push an innocent person onto the tracks below in front of the runaway trolley, where that one innocent person's mass will bring the runaway trolley car to a stop before it hits the five innocent people further down the track.

Typically the majority of respondents to the two scenarios will deem the lever case to be a morally acceptable case of sacrificing one-for-five, whereas not so in the footbridge case. This presents a puzzle to some ethicists: Why are the moral judgments different when the situations seem identical in their morally relevant features? Indeed, some critics suggest that this apparent disparity exposes the deficiency of relying upon moral intuitions for moral guidance.[327]

I argue, however, that the unwarranted disparity is *only* an apparent one: upon deeper examination, it becomes clear that morally relevant features in the two cases are not identical. To see this, we must understand that moral intuitions are not just reactions to singular actions, but also to moral character dispositions as well. This nuanced understanding of what moral intuitions are about is essential to intuition credibility determination: if the moral intuitions in the trolley case are understood as resonant when directed toward the appropriate object of moral character, they should be taken as credible; if they are misunderstood as directed only toward singular agential actions, they

[326] Originally presented by Phillipa Foot (1967).

[327] For instance, Jonathan Haidt (2001; 2003) seems to suggest via the social-intuitionist model, that moral intuitions are largely affected by emotions and are often unreliable. He subscribes to an error-theory of morality, which treats moral judgments like artifacts of causal forces, which explain their normative flavor.

must be deemed noncredible, because they would then "about" the wrong object, in being incomplete. This would signify lack of conceptual clarity, which is an error-disposed condition: the moral intuition is misattributed to the wrong object, or the object itself is not sufficiently understood.

If we understand moral intuitions correctly as arising from judgments of agential moral character, I will show that dispositional utilitarianism concurs with our moral intuitions regarding what the agent in each of the two trolley versions ought to do. This dispositional approach, then, reconciles the ostensibly inexplicable disparity between two purportedly identical cases – flipping-the-switch versus pushing the fat man – a disparity that some critics have used as evidence against moral accounts that rely in some way upon moral intuitions to provide moral guidance.

R. M. Hare (1981, p. 139) derides hypothetical trolley examples – which are often used to test utilitarianism intuitively – as a ridiculous philosophical obsession of "playing trains." Despite this dismissal, the trolley thought experiment, among others, is the subject of the burgeoning field of "experimental philosophy." As previously discussed in chapter 1, a recent study by Greene (2002) utilizing fMRI brain-scans suggests that when people imagine themselves in a situation where they are in close contact with a person, such as pushing a fat man off a footbridge and onto the trolley tracks below, they will not decide to sacrifice the person for the (utilitarian) greater good (p. 178).[328] However, when they do not imagine themselves to be in close-contact with the person, but somewhat remote where their action involves flipping a switch rather than pushing the person, research shows that they will make the sacrifice without much tortured moral deliberation.[329] The resulting data sets are interesting because the two scenarios – pushing the fat man versus flipping the switch to stop the runaway trolley car – seem identical in all morally relevant respects. Peter Singer suggests this tension has been used

[328] Greene claims that because people have a robust, negative emotional response to personal violation in the footbridge case, they immediately deem it immoral; contrarily, people tend not to have a strong emotional response in the relatively impersonal case, and therefore revert to the obvious "minimize harm" principle.

[329] This applies as well in the case were the two cases are pushing the fat man from the footbridge versus flipping the track-switch, which diverts the trolley to make a loop upon the tracks upon which the fat man is tied, in order to stop the hurtling train before it reaches the five possible victims, also tied up. In both instances, the fat man is used as a means to an end, but only the latter seems intuitively objectionable.

to erode the credibility of moral methodologies that rely upon intuitions (2005, p. 332). Since the two scenarios seem to share identical morally relevant features, how can we explain – much less justify – these two conflicting moral intuitions?[330]

The premise of this question, I believe, is ill-founded: the two scenarios are *not* identical in their morally relevant features. We need to be aware of the 2nd caveat: dispositional sets. The two trolley situations are identical insofar as they concern the agent making sacrificial trade-offs: killing one to save five. However, the moral intuitions about such examples are based upon more than just consequences: they also are based upon dispositions of character.[331] For instance, we might tend to morally evaluate an agent who could flip a switch to kill one to save five as having a moral character consistent with being morally upright; contrarily, we would tend to be more hesitant in morally evaluating an agent as morally upright if he could – face-to-face with the victim – push the human being to his death in front of the trolley car.[332]

An agent's actions arise from his/her dispositional set. Actions do not arise *ex nihilo,* and therefore should not be morally assessed in isolation. Some dispositional sets preclude the possibility, for instance, of pushing the fat man. I would speculate that those fMRI subjects who refused the decision of pushing the fat man would also tend to agree with the claim that pushing the fat man in front of the trolley car is not an action open to a person with a morally upright dispositional set.[333]

To clarify this point, let us consider an alternate version of the trolley scenario: A mother, realizing that it will save five lives, can sacrifice her toddler by throwing him under the wheels of an oncoming train.[334] Now consider two different agents: A

---

[330] As indicated by the fMRI research, the emotional centers of the brain light up in the fat man case, but not in the switch-pulling case.

[331] The term "dispositions of character" doesn't necessarily denote a virtue ethics sense of "character," but more generally an agent's disposition to act in certain ways while not in other ways.

[332] Whether or not it is true the agent capable of pushing is actually morally upright, and what determines this, is another question; the point asserted here is that our intuitions take dispositional character into account.

[333] At least not a dispositional set familiar to us as human beings given our societal context: that is, we are not easily familiar with any dispositional set where an agent could push the fat man and still be morally upright (morally upright being, for instance, a person who is disposed to contribute to -- or at least not significantly harm -- social welfare).

[334] In this alternate version, it would seem immoral for the mother to flip the switch as well, but the point being made here isn't in regard to a disparity in pushing versus remote switch-flipping.

sacrificing mother who is easily capable of hurling her child under the wheels of the oncoming train, thus maximizing utility in that instance; and an unwilling mother who is incapable of making such a gruesome sacrifice of her child, even though she recognizes it will save five lives. Our intuitions, I imagine, are that a mother who can readily sacrifice her child in such a way, even for the greater good is a bad mother: she must be emotionally callous in order to be capable of sacrificing her child at a moment's notice in order to serve the impersonal greater good of utility maximization. Our intuitions about such a case extend beyond what *action* an agent should take, but also concern what dispositional set we are to evaluate or endorse the agent to have.[335] In fact, there seems to be a natural association between actions and moral character, as the former serves as indicative of the latter. Take a third type of mother, the conflicted mother, who realizes that she will save five lives by sacrificing her own baby and who, with heavy heart and pained horror, does make this sacrifice. Our intuitions would judge this mother's action as better than the sacrificing mother's action: however, they are the same action! This evinces that our intuitions resonate from dispositions of character, overflowing into actions, rather than just from the actions in isolation.

If scenarios, such as the trolley scenario, are viewed as pertaining to dispositional sets of character, rather than viewed as merely isolated actions, then the alleged conflict between two intuition prescriptions dissolves, and the two scenarios are not identical after all: they differ in their morally relevant features. This inclusion of dispositional sets as morally relevant category can result in a reconciliation between utilitarianism's purported prescription in the footbridge version of the trolley scenario and our intuition that pushing the fat man is morally prohibited: it's not clear that dispositional utilitarianism would actually endorse that the agent to be of the sort who would be capable of pushing the fat man in front of the trolley versus merely flipping a switch.[336] And even if utilitarianism did endorse this action, it would likely endorse that the agent

[335] By "endorse" I mean to signify moral approbation in addition to a more inert moral evaluation.

[336] The fat man example doesn't seem that troubling to utilitarianism in the first place, as many subjects evaluate pushing the person in front of the trolley as the moral action (though this decision does take longer to arrive at, presumably because the scenario presents features that are emotionally salient).

only push with some reluctance, rather than "easily" or "readily" as in the case of the sacrificing mother, above.

We must then understand moral intuitions as sensitive to moral features in a broader sense: not merely in terms of possible action, but possible dispositional sets that can plausibly lead to an action. It may be the case that a particular action maximizes utility in a certain instance: nevertheless, the morally upright disposition that would lead an agent not to perform the action that would maximize utility may be preferable, in a long-term utilitarian sense.[337] To insist, then, that utilitarianism would endorse the mother to hurl her child under the wheels of an oncoming trolley, for instance, is a false assertion, since performing the action might require the mother to be or become the sort of agent who does not maximize utility over her lifetime.

Revising Traditional Utilitarianism

The strategy of such objections against utilitarianism is to present what action utilitarianism presumably would prescribe, appeal to our intuitions which seem to denounce the utilitarian-prescribed action is wrong, and thereby discredit utilitarianism as incorrectly prescriptive. The response I am providing here is that utilitarianism, once it takes into account the importance of dispositional sets, would deny that it would prescribe the actions it is presumed to prescribe. The objector might maintain that act utilitarianism would in fact endorse such sacrifices: for instance, the mother hurling her child under wheel. I would concede that act utilitarianism, understood in a simple way, would deem such sacrifice as the "right action." I would argue, however, that the theory of utilitarianism needs to be expanded and refined in deference to certain background theories, such as human psychology, which would understand agential actions in a more holistic way: not as isolated occurrences, but as manifestations of character.

I contend that this adjustment is not an *ad hoc* adjustment, or at least not an illegitimate one, but a natural adjustment given acknowledgment of relevant factors, and

[337] For instance, utilitarianism could prefer the caring mother who is incapable or unwilling to sacrifice her child over the mother who is capable and willing: the former will have the disposition more likely to maximize utility in the long-run.

their inclusion in theory construction and revision.[338] Consider an analogy to art history, relating to cathedral architecture. An art historian can theorize regarding what aesthetic fashions led to the design of a particular cathedral, but the historian would be remiss not to include also some significant mind to engineering: namely, what internal structures are necessary for the construction of a cathedral? The inclusion of engineering constraints in the art history explanation of cathedral architecture is a natural inclusion. In a similar way, human actions might be considered the metaphorical architecture, whereas the constraints of human psychology are the internal edifice, which is necessary for us to understand when assessing the possibility of actions.[339]

Act utilitarianism, traditionally understood, prescribes that an agent perform that action reasonably expected (subjectively or objectively) to maximize utility. The consequences of the action carry the utility, and so the resultant consequences factor into what action is to be endorsed. Taking a dispositional utilitarian view, in contrast, includes into the calculus the change of character that would result in the mother, were she to sacrifice her baby – even if this would be the "right action" in the act utilitarian sense. Dispositional utilitarianism, inclusive of human psychology, would presumably expect this mother's "right action" to likely result in future consequences that would bear net negative utility: for instance, that the mother's grief, and other psychological effects, would devastate not only the mother's own happiness, but would undermine her potential as a utility-maximizer in the future.[340] Dispositional utilitarianism, then, includes consequential changes in dispositional sets that can be reasonably expected to affect utility-generation in future decisions. If utility is what matters, according to utilitarian theories, then utility-production not just in a singular instance, but over an agent's

[338] Just as a retributive theories might be neglectful of not taking dispositional sets into account in regard to severity of punishment, where an action must be seen within the context of the person's disposition.

[339] I nicked this analogy from Daniel Dennett's (1996, p. 217), though his use of the analogy seeks to communicate the constraints of the evolutionary design space in natural selection.

[340] In illustration of this idea, we might consider young men of the past who have fought in wars, killing enemy combatants. Presume, for the sake of argument, these individuals were acting to maximize the good: i.e., killing enemies that threatened civilization. War veterans, involved in such conflicts, have reported that once you cross that barrier of killing another human being, and it even becomes common, that you cannot readjust to a peacetime society. The psychological effects of war are not entirely well-understood, but are certainly well-known.

lifetime, should be taken into account by any utilitarian theory worth its salt. Act utilitarianism, as traditionally understood, should then be rejected as overly simplistic if it does not incorporate relevant factors, such as human psychology relating to dispositions to act: after all, human psychology does bear upon future utility-production.[341]

As I have been arguing, human psychological dispositions should be naturally included in utilitarian theory: even act utilitarianism itself. Bentham, for instance, first formulates act utilitarianism to take into account the likelihood taking a certain action will lead to future utility-producing actions. Bentham includes this property in the "felicific calculus," terming it "fecundity": the likelihood the action will lead to additional pleasures beyond the initial ones. Bentham, however, locates fecundity in the pleasures themselves rather than in any agent experiencing the pleasures. It would seem a natural extension of Bentham's idea, however, that we expand fecundity to include the agent's disposition toward generating utility in the future. Take the following example: An adolescent basketball player, during the last seconds of a playoff game, considers whether to pass the ball to a teammate or shoot the ball herself at the top of the key. If an adolescent basketball player passes the ball to a teammate underneath the opponent's hoop, rather than shooting it herself, she will get the immediate pleasure of her team scoring an easy point, and the pleasure of her assist. This pleasure is fecund in its leading to later pleasures in locker room, where she is lauded by her teammates. Furthermore, from her teammates' appreciation of her teamwork, she may also cultivate a reinforced and more robust disposition to further cooperate in future contexts where it maximizes utility. During this playoff game, then, she ought to pass the ball to her teammate rather than to shoot, herself. If we were to presume that the pleasure calculus of passing and shooting actually matched up (shooting at the top of the key carries a high pleasure payoff, but a lower probability of success) – even in terms of Bentham's fecundity (possibly equal accolades in the locker room) – it would seem our expanded notion of

[341] If instead of human beings, utilitarianism dealt with programmable robots agents, then act utilitarianism, with its narrow-oriented definition of "right action" might be more appropriate. But ought implies can: and a mother, as a human agent, cannot readily sacrifice her child to the freak-occurrence of a runaway trolley car while concurrently being a loving mother.

fecundity relating to dispositions would favor passing rather than shooting. That is, in the future, this fecundity is carried not in a pleasure's disposition to lead to future pleasures, but in the agent's disposition to act in such a way that leads to future pleasures. In the felicific calculus, this seems a natural inclusion under Bentham's fecundity: the likelihood that a pleasure will cultivate dispositions that will produce future pleasures.

Dispositional utilitarianism extends further than this, however, as I will explore later: namely, it includes the prescription that dispositions need to be developed, and ought to be morally assessed, rather than just actions in isolation. This revision of act utilitarianism to a more robust theory –dispositional utilitarianism – matches up with the reality of how we, in fact, morally judge: that is, we appear to morally evaluate not just actions but character dispositions in terms of their likelihood to lead to certain actions. In this way, the moral theory of dispositional utilitarianism more fully incorporates how the world actually works (e.g., how our moral judgment actually functions) as empirically-related theories ought to do.

In illustration of dispositional utilitarianism's incorporation of psychology in its moral prescriptions, consider a relatively simple case of a man and his dog. Imagine that a man arrives home to see that his dog has knocked over and rummaged through the kitchen trashcan. To make the case simpler, imagine that this man is a utilitarian who brackets aside notions of blame, so whether or not the dog is deserving of punishment is not at issue for the man: he simply wants to maximize future utility for himself and for his dog. The man, upon seeing the dispersed detritus, considers two possible courses of action. On one hand, he knows he could beat his dog harshly with a newspaper, which would thereby ensure the dog never knock over the trashcan again, thereby increasing utility (assume that the human's happiness of a tidy household outweighs the dog's happiness of rummaging through trash in addition to the initial harsh beating). On the other hand, the man also knows that if he beats his dog harshly, this act will change the dog's disposition from an affectionate, happy dog into a fearful and cowering one. Not only will this change in the dog's disposition be a loss of utility to itself over its lifetime, but – to put that aside – the change in the dog's disposition will lower the happiness the owner enjoys from his congenial relationship to his dog, where his dog greets him at the

door with unbridled affection. Aware that the disciplinary beating will change his dog's disposition in this way, the man, as a utilitarian, would calculate that he would maximize utility by granting amnesty to his dog, even though – ideally – he would like his dog to be both happily affectionate and well-behaved. He recognizes, however, that it is not possible that his dog possess both characteristics, given canine psychology (or at least the psychology of his particular dog).

A similar disposition-based story can be related to human beings and ground-projects. Utilitarianism would only demand sacrificing one's ground projects if this abandonment would result in greater utility in the long run than if the agent were not divorced from such projects. Ground projects, we must recognize, imbue the agent's life with meaning. If giving up such projects were to destroy personal motivation, hollow out an agent's enjoyment of life, and remove incentive to act positively, then utilitarianism wouldn't really endorse sacrificing one's ground projects. To require such sacrifice would be to nullify an agent's future positive actions. After all, we cannot expect human beings who would maximize utility in every instance to be actually possible, sustainable, or desirable. Perhaps androids could be programmed to maximize utility (given a lithium battery in place of ground-projects). But humans are humans, and their psychology needs to taken into account. Relating back to the example of the mother sacrificing her child, act utilitarianism, if simply understood, might prescribe she perform the "right action" of sacrificing her baby. A more sophisticated version of utilitarianism would not prescribe this action, however, as this sacrificing could be reasonably expected to result in lesser utility production over her lifetime. The objection, then, would be mistaken in asserting what a more sophisticated version of utilitarian would prescribe. Consider that even act utilitarianism might not prescribe the mother sacrificing her baby to the runaway trolley car if the mother, wracked with monstrous grief and guilt, would with a high likelihood commit suicide later that day.[342] If the mother does not commit suicide, but instead

[342] Let us just imagine that this loss of life, though, granted, it's unlikely to work out this way, causes grief that outweighs however many innocent persons were on the trolley tracks, and that we are calculating utility in an aggregative way, where her loss of life would suddenly remove any positive net utility that she would experience over her lifetime.

becomes, a hollowed-out shell of a person, dispositionally incapable of generating utility for herself or others in the future, doesn't that essentially amount to the same thing?

Mothers and Robots: Utility and Time

Of course it could be the case that even if the mother sacrifices her baby, and becomes a hollowed-out shell, incapable of generating utility – that her saving the five lives of the innocent people outweighs any utility she would have otherwise generated, even over a lifetime. In such a case, would not dispositional utilitarianism prescribe that she sacrifice her baby? The short answer is no. Dispositions come with probabilities that a certain character trait set will maximize utility over a lifetime. Dispositional utilitarianism endorses those dispositions that can be reasonably expected to maximize utility over a lifetime, not that actually *do* happen to maximize utility over a lifetime (given some unforeseen occurrence). After all, we cannot account for the happenstances of runaway trolleys, hobo organ transplants, airplane crashes involving world-renowned surgeons, and the like. Furthermore, we should not cultivate dispositions to be ready to maximize utility in such situations. Having a disposition to maximize utility in such bizarre situations would not be coherent with dispositions to maximize utility ordinarily.

In illustration of this tricky point, I request the reader's indulgence in the following analogy: compare the programming of robots with varying task-strategies, to the cultivation of human psychology with varying dispositions. Consider the case of a simple walking robot. Imagine that the walking robot is blind to its environment, but only has the simple task of walking from point A to point B as time-efficiently as possible. Typically, a robot can only be given a finite set of strategies or heuristics to reach a certain goal. For instance, for our walking robot to get from point A to point B – across a room – in the shortest time, we would program it to walk in a straight line until it reaches point B. Imagine, however, that in the room environment, there are pillars as occasional obstacles. Then we might program a robot to walk in a straight line until it encounters an obstacle, in which case it should employ the following strategy: turn 90 degrees and take a step, and turn back the 90 degrees, and try again to step forward – repeating this

strategy until the robot could resume forward movement. This “turning” robot would have a better chance of reaching point B in the shortest amount of time than the “straight-line” robot that repeatedly walks into a pillar without any program to overcome this obstacle.

It might happen to be the case that room is an “ordinary world” where nearly all obstacles are pillars, and so that the turning strategy is the optimal strategy for maximizing time efficiency. But suppose it turns out that, in one bizarrely rare instance, a 2 inch by 10 foot wooden-plank presents itself as an obstacle. A robot that had the strategy to “step-over” the plank would maximize time in this singular instance. Given this, we might consider adding the “step-over” strategy into our turning-robot’s repertoire. If we did this, the turning-robot would now employ the stepping-over strategy first when it encountered obstacles, and if meeting no success, would secondarily employ its turning strategy. However, given our ordinary world where this wooden-plank obstacle is so rare, it really would be wasteful to reprogram the turning-robot in such a way. In fact, if we did so program the turning-robot with the “stepping over” function, it would generally have much worse time-efficiencies than the robots that just had the “turn” function.

In a similar way, mothers (like the turn-robots), could be cultivated to sacrifice their babies in bizarre trolley car situations (like the stepping-over robots), if such situations were ever to arise, but if mothers were so cultivated, they would not typically maximize utility over a lifetime (in a similar way that stepping-over robots would not maximize time-efficiency over a lifetime of trials in the ordinary environment). Imagine how the dispositions of mothers would have to be if they were to be ready to make such split-second sacrifices: they certainly could not value their child so disproportionately in relation to others. But we recognize that this maternally biased-love of a mother toward her baby is what benefits the mother, the child, and our community. Furthermore, human beings are not as psychologically malleable as robots, where what we often characterize as “unconditional maternal” love could countenance numerous escape clauses or

exemptions.[343] For a mother to be capable of escape clauses, she would seem to need to suffer dissociative identity disorder, and have at least two distinct personalities: one, a mother of unconditional love, and the other more akin to Euripides' Medea.[344]

If it were the case that trolley-car situations occurred more frequently, just as in the case where wooden-planks frequently occurred as obstacles, then such cultivation (programming) would lead to maximized utility (time-efficiency). I will explore this in the next section, considering mothers in the Inuit culture who not uncommonly must sacrifice her baby for the greater good. Therein, I discuss moral function and dispositions. I will argue that utility dispositions relate to moral praise and blame. For instance, in our mother-trolley case, it seems we should not expect the mother to be able to readily sacrifice her child, having a disposition of unconditional maternal love, which she presumably cultivated in preference to viewing her baby as expendable if the price-is-right. And, relatedly, it seems we should not blame her for lacking such wild escape clauses.

In summary, our intuitions about the footbridge version of the trolley case are said to be at odds with utilitarian prescriptions. However, the respective intuitions do not have as their objects just singular actions, but also the dispositional set of the agent. The prescription of our moral judgment, upon deeper examination, may not simply be "Do not push" but may rather be the broader prescription "Do not be or become the sort of person who pushes, and so do not push."[345].

---

[343] That said, we must recognize that there are some "escape clauses" in human morality, though limited in nature and scope. In scope, it appears that numbers do matter if the numbers are large. Even the deontological condemnation of torture intuitively allows the exemption of the "ticking time-bomb" case, where torturing a person would be morally permissible to save the greater number from peril. Moral permissibility seems to bleed into moral obligation in cases where number are vastly increased, such as where the agent must kill an innocent bystander in order to save the entire European Union from nuclear annihilation. Refusal in such cases might strike us as a bit precious, and even our solid-as-oak deontological convictions of "never sacrifice an innocent for the greater good" might falter. In addition to number, it appears that the nature of our morally-pertinent landscape matters to our evaluation of moral dispositions. Killing in a time of war, for instance, is more acceptable than killing in a time of peace.

[344] *Medea* is a tragedy written by Euripides, first produced in 431 BCE, in which Medea kills her two sons, whom she allegedly loves, in order to avenge the betrayer of her lover, Jason, who is their father.

[345] There is the further worry that if you push the fat man, you will *become* the sort of person who pushes: a person who does not maximize the good (not necessarily a utilitarian good, but good in a common, general sense).

If these intuitions are viewed in this broader way – that they are intuitions about dispositions rather than mere singular actions – then the intuitions do not necessarily conflict with utilitarianism; utilitarianism can endorse acting from the disposition that makes pushing the fat man an action not capable by the agent, while she still has a utility-optimal dispositional set. After all, utilitarianism cannot be presumed to prescribe a particular action without first knowing that the action is in fact open to the agent.[346] Utilitarianism must inquire as to what an agent's dispositional character must be like in order for the agent to be capable of committing certain actions. We cannot expect utilitarianism to endorse, for instance, a mother who can, at any sudden moment if greater utility demands it, boil her baby.[347] In order for this action to be open to the mother, she needs to have already a certain dispositional set in place. I don't believe it's specious to suggest that a mother who is readily able to boil her baby for the greater utility is both a bad mother and a bad human being, both in the evaluation of our commonsense morality, as well as concerning dispositional character relating to utility-maximization. We should not let our imaginations promiscuously conceive that it is possible that caring mothers can simultaneously be utility-maximizers without constraints.

States of the World and Right Action

In order to elucidate fully the notion of dispositional sets, and our moral judgments related to them, I will explore and distinguish, in the next few sections, five

[346] This is not to excuse bad behavior. A person who acts badly from a bad disposition should strive to improve their disposition (though they might not see it in terms of utility): presumably through good actions. And it does seem like a bad disposition serves as at least some excuse for suboptimal behavior: we would tend to evaluate as morally superior, for instance, the achievements of a former gang-member over the son of a senator. Even if the Senator's son does more good, we presume he had less obstacles to overcome, and that his moral disposition was cultivated more from benevolent external factors (and an absence of malevolent ones) than from an internal struggle with demons within.

[347] The example of boiling a baby can be originally attributed to G. E. M. Anscombe (1981), who used this example in a 1956 pamphlet that she wrote and distributed, which protested the conferral of an honorary degree to President Truman by Oxford; her protest was largely based on her condemnation of President Truman's decision to drop the atomic bomb on Hiroshima and Nagasaki.

concepts: (1) states of the world (2) right action (3) dispositional sets (4) praise/blameworthiness and (5) moral function. In order to understand what our moral judgments are judging, we need to analyze these five concepts, and be cognizant of their relations and distinctions amongst each other. This greater resolution will help us perceive the crucial nuances that are at play in anti-utilitarianism thought-experiments.

The first concept above is also the simplest: states of the world. This term refers to the possibilities that can result from various actions an agent selects. Generally, act utilitarianism prescribes an agent perform the action, among a set of possibilities, which produces the state of the world where utility is maximized.

Act utilitarians use the term "right action" to refer to that action that maximizes utility: that is, the action that achieves the state of the world with optimal utility in comparison to other possible states. I will redefine the term "right action," later, in relation to dispositional utilitarianism. I admit some hesitancy in employing the term "right action" as I suspect it might court confusion: a "right action" denotes that action that act utilitarianism would prescribe an agent perform, which is presumed to be that action which results in the utility-optimized state of the world in that instance. Yet "right action" seems colloquially to suggest that the agent, herself, performed the action that she should have. I do not believe these two concepts need necessarily be married. An agent can "fail" to maximize utility in a particular instance, and still have performed that action she should have performed. Similarly, an agent can maximize utility and yet not have performed that action that she ought to have. I believe this is consistent within a utilitarian paradigm, where this "ought," in both cases, needn't be cashed out in terms of moral value beyond utility. For instance, given dispositional utilitarianism, an agent can fail to maximize utility in a particular instance because to perform the action was not open to her dispositional set, or would have altered her dispositional set to a sub-optimal set.

In this chapter so far, I have been arguing that utilitarianism prescribes optimizing utility in the long-run, which is dependent upon the dispositional set of a given agent: this dispositional set precludes, oftentimes, singular actions which would result in an optimal state of the world in a particular circumstance (e.g., a mother boiling her baby to save ten

lives). “Right action,” then, needs to be distinguished from states of the world; the term “right action” should only be used to denote the action that utilitarianism would prescribe in deference to dispositional sets, where a “right action” must be seen in the context of an agent’s propensity to maximize utility over a lifetime.

Under the paradigm of dispositional utilitarianism, we might redefine “right action” the following way: An action is right if and only if it maintains or cultivates a dispositional set that could be reasonably expected to maximize utility over the course of the person’s life when compared with acting from some other dispositional set. I will refer to this definition as the first definition.

An alternate, second definition of right action for dispositional utilitarianism might be expressed in the following way: An action is right if and only if it would be done by a person with a dispositional set such that acting from that set could be reasonably expected to maximize utility over the course of the person’s life when compared with acting from some other dispositional set. I argue that this alternate definition is not preferable to the initial definition provided above. There of course could be a reasonable debate over which definition is more sensible, a debate which parallels one in virtue theory.[348]

In argument for the first definition of right action, according to dispositional utilitarianism, let us first consider an example. Imagine a miserly man, Malcolm, who

[348] In virtue theory, the question can be phrased the following way: Is the right action the action a virtuous person would do? Or is a right action the action a person should perform in order to cultivate a virtuous disposition? Robert Johnson (2003), where he argues that the right action is not necessarily the action a virtuous person would perform: sometimes the right action is that action that would lead the non-virtuous person to become virtuous through processes in which a virtuous person would need not to engage. An example Johnson presents is that of a non-virtuous person who is prone to lying. This person might write down his lies, to keep track of them; he would also consider what effects telling the truth would have; and finally he might try to engage in activities that would help increase his self-esteem, which seems to be a cause of his tendency toward lying. Johnson states that none of these activities are ones that a virtuous person would engage in, as they would have no need to do so. I am inclined to agree with Johnson’s argument: the right action is that action which would cultivate the appropriate disposition. One could argue that by performing those actions a virtuous person would do, that a non-virtuous person would eventually become virtuous. It might be argued, however, that a non-virtuous person might not be able to perform those actions a virtuous person would perform, or that the non-virtuous person could become virtuous more quickly and efficiently if he were to pursue means that a virtuous person would not pursue, as they would have no need. Lastly, it seems that in the case of Malcolm, an example in the paragraph following, that his initially performing the right action – as defined as the action a virtuous person would do – is actually counterproductive to his becoming virtuous.

realizes that he ought to give some of his wealth to the needy (perhaps he's convinced by the logic of Peter Singer's 1993 argument in "Rich and Poor"). Suppose that Malcolm realizes that the virtuous person would give to the point of comparable moral significance, where he gives up all of his luxuries, but stops short of sacrificing things that significantly contribute to his well-being and life's meaning. Consequently, Malcolm makes some calculations relating to his finances and life-projects, and realizes that this means he should give up 70 percent of his wealth. With a deep breath, he divests himself of this wealth. Malcolm, having a miserly disposition, reels from his sudden "impoverishment" and vows never to give to charity again, be damned his moral obligations. The second definition of right action would prescribe that Malcolm perform this action of making a 70 percent one-time donation (as this is the action that would be performed by an individual with a dispositional set reasonably expected to maximize utility, compared to other available dispositional sets) even though his making this one-time donation precludes his making any future charitable donations.

Imagine an alternative course of action for Malcolm, where rather than sacrificing 70 percent of his items, he eases himself into it: the first year giving up 10 percent of his income; 20 percent the second year; and so on until he achieves the 70 percent sacrifice after seven years. Over Malcolm's lifetime, this would maximize utility (via charitable donations), whereas the one-time 70 percent donation would not. This is the primary reason why I assert the first definition is superior, in regard to utility-maximization, to the second definition of right action. One possible way of perhaps merging the two definitions is saying that Malcolm ought to perform the gradually increasing charitable donations as a means to performing the "right action," as characterized in the second definition. That is to say that by donating gradually rather than all-at-once, Malcolm will eventually attain the utility-optimizing disposition that he otherwise would never be able to attain.

One possible problem with the first definition of right action regards weakness of will. Suppose a person suffers weakness of will, where he realizes what he ought to do, but cannot always manage to live up to this ideal. We must then ask: What ought this person to do, given his weakness of will? According to the first definition, the right

action would be that action that will lead them toward the disposition where they would likely optimize utility over their lifetime. However, perhaps due to this affliction of weakness, they can never achieve utility-optimization: that is, the individual never can achieve that dispositional set that would maximize utility. Would it be wrong, for instance, to say that Malcolm ought to eventually give 70 percent of his wealth to charity if there is no possibility that he will do so, even after several years of gradually increasing giving? Dispositional utilitarianism would prescribe that Malcolm work to acquire those dispositions that he actually could acquire, given his weakness of will. This would allow him to produce more utility over his lifetime than if he hadn't acquired these dispositions, even if these dispositions ultimately remain sub-optimal regarding utility-maximization.

This prescription by dispositional utilitarianism, however, raises the objection that individuals with weakness of will, like Malcolm, would have a lowered "moral bar" in comparison to others. The right action for Malcolm, for instance, might be to cultivate the disposition to donate 50 percent of his wealth, while others were donating the full 70 percent. The "right action" for Malcolm, then would be a gradual increase to 50 percent, whereas the "right action" for those not suffering weakness of would be 70 percent. Does this not give Malcolm a "free pass," while evaluating him as equal to the person donating 70 percent?

One possible response to this objection is just to accept that Malcolm is performing the right action in the same way his more charitable peers are; the right action could be taken as agent-relative: performing that action among those open to the individual agent that would lead to a utility-maximizing disposition that the agent could possibly achieve. We seem to recognize that people have varying dispositions, and – at times and to some degree – morally evaluate agents accordingly. A man who grew up a poor orphan in a state-of-nature kind of environment, for instance, might have more difficulty giving up his wealth than a man who grew up in an affluent family. Even if the two men had the same amount of wealth, and the adult orphan gave less of his disposable income than the affluent adult, we might still view their giving as morally equivalent (at least in regard to their praiseworthiness). I do recognize that this is a contentious

assertion; I would maintain, nonetheless, that dispositions do seem to matter in moral evaluation. In our judicial system, for instance, character disposition may, at times, constitute a mitigating condition when determining criminal responsibility. For example, an adult child who intentionally kills his parent in a fight might be given less time if that adult child's childhood was rife with physical abuse caused by that parent. Such recognition of mitigating conditions, relating to dispositions, does show that we accept that the "moral bar" for individuals can vary in some circumstances.

A more satisfying response to the objection, however, might be that cases of weakness of will should be treated as so-called "contrary to duty imperatives." These are cases where a person has failed to do what he ought to do, and so the question becomes what the person should now do in light of this moral failure. A person who has lied, for instance, should inform the person that he has lied, thereby empowering the person to prevent or rectify any possible damage that might have been caused by that lie; or the liar ought to make amends for any resulting damages to the person to whom he lied. A person who breaks promises might let his associates know that he has a problem keeping promises, so that they should be wary in accepting promises from him; or he might pay for some of the damages resultant from his broken promise. In the case of Malcolm, he might acknowledge to himself that he will never reach the 70 percent donation ideal, and he might request that others donate monies in his name, rather than giving him gifts on his birthday or on holidays. Such actions are not those that optimally-disposed agents would do, as such agents would not be in a position where they would need to do so. But these actions are actions that the sub-optimally-disposed agent ought to do, given that they will not have done what they ought to have done in the first place.

In this way, we can evaluate Malcolm as morally inferior to, say, Peter, who is able to donate the full 70 percent to charity. Malcolm, then, did not perform the "right action," but performed the second-place right action, given than he failed to perform the first; by donating 50 percent, he has done the "second best" thing. We might further blame Malcolm for not developing utility-maximizing dispositions earlier in his life, when such cultivation was possible; of course, it might not have been his fault, at least not entirely, for the dispositions he finds himself with later in life. We can maintain that

Malcolm has the same moral bar as others, but that he just will consistently fail to reach that bar. Nevertheless, knowing in advance that he will fail to reach the utility optimization that is typically possible by human beings in a certain environment, we can indicate what the next best thing is for him to do. I would maintain that this "next best thing" might be appropriately deemed "the right action" relative to that sub-optimizing agent, as the next best thing is the only action open to him, given his dispositional set, and it is therefore the action that he ought perform, given that ought implies can. Nevertheless, Malcolm can be assessed as morally inferior to others whose dispositions allow them to produce more utility over their lifetimes than Malcolm.

Dispositions and Praise/Blameworthiness

As already discussed, dispositional sets, refer to the propensities of an agent that dispose him toward certain types of actions. An action is possible only if an agent has a compatible dispositional set. For instance, a loving mother, as previously discussed, cannot readily throw her infant under the wheels of an oncoming trolley car (even if she knew it would maximize the state of the world in that instance). If the mother were easily capable of doing so, she would not have the dispositional set recognizable as that of a "loving mother." Though this might sound circular, it seems rather a matter for empirical psychology; for now, I assert it *prima facie*. If it did turn out that a loving mother could make such a macabre sacrifice, then utilitarianism would prescribe that mothers become such a person; I, however, think that such a possibility is unlikely.[349] It should be noted that actions are to be morally assessed not merely in terms of the disposition the agent actually has, but in terms of the disposition the agent could be reasonably expected to have, given the environment – involving both internal and external factors – in which her disposition was cultivated. An agent is responsible for not only failing to choose the utility-maximizing option open to their dispositional set, but failing to cultivate the

[349] As I already stated, much of the "possibility space" for dispositional sets is dependent on psychological research. I do fully endorse appealing to empirical evidence in moral psychology; however, I believe the claims I have been making, such as concerning a loving mother, will be accepted as relatively uncontroversial. For that reason, I will leave these assumptions empirically unanchored.

utility-maximizing disposition that would open to possibility further utility-maximizing actions. That is why it would be wrong for a person to defend their bad actions with the justification, "That's just how I am!" The obvious response to this is, "Well then, you should work to change how you are." This change in disposition, however, needs to be both possible to the agent, and a change we can reasonably expect them to achieve. An only child cannot be expected to readily share their toys with other children upon their first associating with other children, but this disposition needs to be cultivated. Likewise, we cannot expect a novice police officer to be as easily able to shoot an armed robber as a seasoned officer with years of in-the-field training.

Praiseworthiness and blameworthiness builds upon dispositional sets: it regards whether or not an agent is deserving of moral commendation or condemnation for her action or dispositional set. Praise/blameworthiness typically has a close connection with a person's dispositional set, but neither is necessary for the other: that is, one can be praised for a bad dispositional set, or blamed for a good dispositional set. Praise/blameworthiness, I will assert, is another way of saying that an agent is functioning well, in a moral sense: this introduces the fifth concept of "moral function." An agent, in rare cases to be discussed, can be morally functioning well, but have an unproductive dispositional set: a set that does not maximize utility. By "moral function," I do not mean to introduce an altogether new theoretical concept, but rather to capture what I believe we mean when we deem an agent to be "praiseworthy:" the agent is thought to be praiseworthy *because* he is thought to be functioning well, morally. To clarify these concepts, and to disambiguate dispositional sets from praise/blameworthiness and moral function, I will explore parallels to evolutionary biology.

Parallels to Evolutionary Biology: Fitness and Function

Biological fitness parallels dispositional sets, and evolutionary function parallels praise/blameworthiness. Fitness and dispositional sets are more descriptive than

normative:[350] they concern the relation between an individual's propensities given its environment. Evolutionary function and praise/blameworthiness, on the other hand, are more normative than merely descriptive: they concern how an individual is *supposed* to be, given a certain background.[351] This background is the moral upbringing of the agent, both culturally and evolutionarily. I will not delve into the concept of "background" further here, as I believe it will become clear via the examples that follow.

Fitness, as characterized by evolutionary biology, is the degree to which an individual's features increase their chances of propagating genes into future generations. Similar to this, dispositional sets with which utilitarianism as we understand it is concerned are those features of an individual that disposes the individual to optimize utility, overall. In this way, we might even call an individual "morally fit" if he possesses a dispositional set that can be reasonably expected to maximize utility over his lifetime in a certain environment. In both the moral and biological dimension, an individual's dispositional set – whether in regard to biological fitness or agential action – depends on the environment. Fitness, for instance, is a quality that supervenes upon an individual's features in relation to the specific environment in which one is situated. Likewise, a dispositional set is a quality that supervenes upon an individual's features in relation to the specific environment in which the individual is situated.

Function is different from fitness, in biology, in that an animal can be functioning well, yet could be unfit. Function and fitness are usually highly correlated, however, insofar as functions evolve in order to increase fitness: for example, a cheetah evolved its speed in order to genetically survive.[352] At times, in nature, fitness abandons function. Global warming, for instance, has caused some perfectly well-functioning migratory

---

[350] Dispositional sets can be considered as prescriptive in that dispositional utilitarianism prescribes those dispositions an agent ought to cultivate: namely, a productive dispositional set. By "descriptive," here, I mean that there is nothing intrinsically normative about the concept itself. Similarly, "utility" is a descriptive concept (e.g., happiness), whereas "praiseworthiness" is a prescriptive concept, where intrinsic to the definition is that deeming an agent praiseworthy is to say the agent *ought* to be praised.

[351] What I mean by "supposed to be" is how onlookers would judge how the agent should be. This comes into folk moral judgments regarding moral functioning; I am not seeking to justify the concept here, but merely to describe it in order to arrive at some clarity between the five concepts.

[352] The teleology here is palpable, but teleology is accepted in biological fields. We can rephrase our statements to drop out the teleology: i.e., ostensible "function" exists because it was the most fit of the alternatives that happened to have come up.

birds to migrate seasonally off-schedule, much to their peril. Though these birds are functioning well, they are no longer "fit" due to drastic environmental change.[353]

Biological function, in contrast to fitness, is treated as normative concept: the heart functions well if it pumps blood (rather than makes hearts sounds, though it does both).[354] Fitness, however, is merely a descriptive concept (in the same way as are dispositional sets): something is fit if it has a certain relationship with its environment, which enables it to propagate. If an individual is functioning well, it is highly likely the individual is also fit, and vice versa. In evolutionary biology, an animal can be fit but not functioning well: for example, a random genetic mutation of a particular finch on a Galapagos Island might cause it to have a stout beak rather than a needle-nose beak, which has traditionally functioned to retrieve insects from the trunks of trees. If the niche of siphoning insects from trees has dried up, however, this mutated stout-beaked finch might be functioning poorly while being optimally fit in relation to the environment – say, if his stout-beak could be serendipitously employed to exploit a new niche, such as cracking open nuts for nourishment.

In the above ways, we can see how fitness and function, while frequently conjoined, are separate and distinct. In the same way, dispositional sets and moral function, which I believe directly underlies praise/blameworthiness, are frequently conjoined, but they are likewise distinct and separable. In regard to utilitarianism, a dispositional set is that set an individual has which disposes her toward generating a certain degree of utility. A friendly neighbor, for instance, has a dispositional set that tends, on the whole, to lead to more utility than that of a self-centered neighbor. We could describe that friendly neighbor as having a more productive dispositional set than the self-centered neighbor, in that it produces greater net utility. We would, likewise, say

[353] Also, an organism can be fit but malfunctioning, such as humans with sickle-cell anemia in malaria-stricken countries. Sickle-cell anemia happens to protect those afflicted against contracting malaria; this dramatically increases their fitness. One might, of course, over a period of successive generations, deem sickle-cell anemia as a "selected-for" trait, if it were genetically-inherited and increased fitness, and call it increasingly appropriate to deem it a function (or in the very least, decreasingly appropriate to deem it a malfunction).

[354] An excellent discussion of function occurs in Larry Wright's 1973 article, which is where I retrieved this example.

she is functioning well, morally, in being somewhat altruistically disposed toward her neighbors.

Again, to clarify, I do not mean "functioning well" as a utilitarian description. In fact, utilitarianism might at first appear to be blind to moral function; utilitarianism only recognizes dispositional sets. I intend "moral function" to be taken in a broad non-theoretical sense, and not tied down to any particular moral conception. I bring in "function" as an explanation of how people morally evaluate themselves and others: simply put, that an agent happens to be intuitively judged as praiseworthy if he is considered to be functioning well, and is judged as blameworthy if he is considered to be functioning poorly. The functioning of an agent is not ahistorical: it is connected to an agent's developmental background.[355] In cases, such as the trolley example, I believe individuals judge the switch-flipper as functioning well (or well enough) whereas the footbridge-pusher is functioning poorly, given the developmental background: that is, that you ought not to sacrifice those in your in-group, according to the agent's culturally and evolutionarily inculcated background.[356] This last evolutionary reason has been

[355] It seems in recognition of developmental background that we feel justified, in certain cases, in dialing down our moral assessment of an agent's blameworthiness if his developmental background makes his moral offense less anomalous: e.g., we might be less condemnatory of an individual who is sexist who comes from a sexist culture than a sexist who comes from a culture where sexism is not instilled. We must be alert that condemnation of the agent is what is being talked about here, not condemnation of the behavior: in both cases, we would condemn the behavior, but the condemnation of the agent himself would differ.

[356] This notion of moral function, given developmental background, brings up the interesting case of moral reformers, such as Huck Finn (previously discussed in Chapter 3). We might ask: Is Huck Finn not functioning well if he is acting in conflict with his developmental background? I believe it could be said that Huck recognized, at least via his emotional sympathies (his implicit recognition of Jim the Slave's humanity) that certain culturally-presumed facts were false, via his association and friendship with Jim. To extend the evolutionary/biological comparison, we might venture to analogize that Huck's moral shift was like an increase in fitness and well-functioning. Huck's moral shift may not have been a change of Huck's moral values, but a revised understanding of the facts and objects to which those moral values applied. Regarding praise/blameworthiness, his society would consider him blameworthy; we, on the contrary, would consider him a moral exemplar, and praiseworthy, akin to many moral reformers. We would consider him functioning morally well, not just because he happens to be right in his moral judgment, but that his moral faculties are functioning so well as to allow him to shift his moral paradigm against the crushing weight of his society's instilled prejudices and presumptions. This observation does not invite claims of relativism – Huck actually is praiseworthy and is functioning well: his society clearly had the facts wrong regarding the subhuman nature of African Americans, which became clear to Huck through his friendship with Jim. In essence, Huck is able to be motivated by morally relevant features, and not overrun by morally irrelevant features (e.g., false facts, confused concepts, distorting emotions, and so forth). Moral function, then, involves an ability to act in deference to morally relevant reasons – that is, the "right

proffered by Singer (2005, pp. 339-342); he appeals to the evolutionarily history of human beings as closely-knit in-groups to explain the disparity between the two trolley-car cases previously mentioned at the start of the chapter.

Actions, Agents & Persons

Before launching into the discussion further, it might be helpful to disambiguate some underlying concepts. There are three kinds of moral evaluations: evaluations of actions, agents, and persons. Actions are to be evaluated as right or wrong. Agents are to be evaluated by their dispositions to produce utility. An agent can be morally good or morally bad; they may have certain character traits such as trustworthiness, dishonesty, selfishness, generosity, cruelty and so forth. An agent is "morally fit" if he is disposed to maximize utility over their lifetime in the environment they are in. By using the term "persons," I am casting the individual as more robust than a mere agent: persons are not to be assessed in mere relation to their dispositions, but in their dispositions in relation to certain backgrounds (a claim which will be developed later).[357] Persons are to be assessed as morally praiseworthy or blameworthy in relation to how they could have been expected to have acted, given their background. In the same way, they can be said to be morally functioning poorly or well.

In elucidation of these concepts, we can consider some examples. Since these three concepts are interrelated, however, it is challenging to generate intuitive examples where the concepts are parsed from one another. A further challenge is that terms like "right action," "bad person," and "blameworthy" are often used colloquially and without

---

reasons." Presumably this is what Jonathan Bennett in his article was proffering: That Huck was functioning well, morally, by having his sympathies and conscience/principles in balance in such a way where his sympathies could check/correct his principles.

[357] Previously, I was using "agents" and "persons" synonymously. Only in this section will I be using agents in a narrow sense where they are stripped of their developmental backgrounds, and only seen in terms of dispositions. One interesting thought-experiment in the philosophy of evolutionary biology parallels this distinction between person and agent, where it distinguishes between an animal and an organic artifact. Imagine that miraculously a collection of atoms coalesced perfectly to form what looked to be in every way a horse. Would it be a horse? Though this spontaneous entity could run fast, could we say it had the *function* of being fleet of foot? The answers to these questions, at least in my opinion, appear to be "no." The ahistorical entity in this thought experiment – void of evolutionary heritage and function – is much like the "agent" used in this section.

philosophical precision. Both of these challenges may cause the examples to be not as intuitively apparent as typical examples might be; for this reason, it may take some imagination on the part of the reader to conceive of the following sophisticated characters as possible in the way described.

First, we might imagine the incapable mother who is committing a wrong action, in the act utilitarian sense, by not sacrificing her baby for the greater good in our aforementioned trolley-car case. But we still might consider her a morally good person: considerate, caring, and so forth. We might further consider her praiseworthy in not being readily able to sacrifice her child to the rare occurrence where it maximizes utility.

Second, consider that a peer who performs the wrong action by frequently lying, and who is morally bad in that he tends to be dishonest: you may want to avoid him for this reason. Nevertheless, you may consider him not to be blameworthy, as you understand his upbringing took place in a notoriously dishonest household. For this reason, you might give him special reprieve in your moral estimation. You might even find him praiseworthy if he is in fact diligently working on changing his disposition (though you still might avoid him until he reaches that goal).[358]

Third, imagine that a mother of a superstitious hunter-gatherer tribe sacrifices her first-born child to the "hippo god" that resides in the river bordering her village. The villagers believe (wrongly) that the hippopotamuses in the river are supernatural entities who demand such sacrifices, or else they will bring divine wrath upon the villagers. The mother would be committing the wrong action, yet could be still evaluated as a morally good person (honest, generous, altruistic); furthermore, she could be deemed praiseworthy (or at least not blameworthy) in that she is making, with much grief and reluctance, this (perceived to be) necessary sacrifice. In this way she is functioning well, morally, despite performing what we (and act utilitarianism – at least a version that conceives of rightness as objectively determined) would recognize as the wrong action.

[358] One real-world example of this might be the example of some of the Vietnamese children saved from war-ravaged Vietnam during Operation Rescue. This rescue operation transplanted Vietnamese children from their home-country to the United States, where they lived with American families who took them in. Many of the Vietnamese children would reportedly chronically steal and lie, as they grew up in conditions where they had to steal and lie to enable the survival of themselves and any younger siblings. This does not seem morally blameworthy, understanding this background.

Fourth, imagine another example of a mother who tragically lives in a runaway-trolley world, where mothers must often sacrifice their child for the greater good. The mothers in this world soon learn to be emotionally callous toward her children in order to be able to make such sacrifices for the clear greater good. Our mother may have been overwhelmed with initial maternal love for her children, but have painfully strived to become deliberately callous toward her children in order to enable herself to sacrifice her children when the greater good clearly demanded. It might be the case that she still struggles to love her children up to a certain point, whereas other mothers – realizing the sacrifices they shall be repeatedly called upon to make – allow themselves to be excessively callous and occasionally cruel in desire to avoid the grief they would otherwise feel upon each necessary sacrifice. Our mother, in these circumstances, might be the best mother possible, in this instance, with the optimal utility-producing disposition. In this environment, she is a "morally good" person in that the term seems to be somewhat comparative: she has the best available dispositional set that a mother can have in this bizarrely tragic trolley-world. She is the most morally fit mother in this environment.

We might whisk this mother away, though, and transplant her into the safe and sunny suburbs of our ordinary world – almost identical to her own yet where these runaway trolley scenarios never actually happen (they only exist in the warped minds of philosophers). The mother would now not have a utility-productive dispositional set, compared to the other ones generally available. In contrast to the other suburban mothers, and their dispositions, her callousness and cruelty toward her children rightly mark her as a morally bad person. In fact, you probably wouldn't want to hire her to baby-sit your kids. Imagine it so transpires that a runaway trolley situation happens in this ordinary world, and the mother is in the thick of it: if she – being so disposed – sacrifices her child for the clear greater good, she would actually be performing the wrong action! An individual with a disposition likely to maximize utility over a lifetime in this environment would not perform this sacrificial action. The community might not only deem this mother to be a morally bad person, but further evaluate her to be morally blameworthy. Nevertheless, as odd as it may sound, this mother, while a morally bad (and unfit) agent

in this environment, is not a morally blameworthy person. She is functioning well, morally, given her developmental background. When we understand the background that cultivated her dispositional set, we recognize that, though as an agent in her current environment she is a morally bad person (cruel and callous), she is still not blameworthy given her background. In fact, we might see her as tragically noble. In this new world, the mother can and should strive to change her dispositional set toward becoming a more nurturing mother – presuming this is psychologically possible for her to make such strides after a lifetime of chronic child-sacrifice.[359]

One last issue should be considered before embarking on the next section: praiseworthiness and blameworthiness in relation to actions versus dispositions. A person can be praiseworthy or blameworthy in relation to either disposition or in relation to action. Consider two negligent doctors who are disposed to prescribe dangerous medication to patients without doing their due diligence in researching the medication first.[360] One doctor happens to have a patient come in asking for the medication, which he negligently prescribes. The other doctor happens to have no such patient who comes in that day. Both are equally negligent in their character, and equally blameworthy in this moral failing of their respective characters. Yet only the first doctor is blameworthy in their action of prescribing the dangerous pills; the other cannot be deemed blameworthy of that since it isn't something that he actually did.

This distinction between actions and dispositions being praise/blameworthy seems intuitive, but there emerges a counterintuitive result in making this distinction: a person can be praiseworthy for doing the act we (and act utilitarianism) would recognize as morally wrong, and blameworthy for an action that we would recognize as morally

[359] A more real-life example of transplantation is the soldier who goes to war. Boot-camp is not just for physical training but emotional training as well: in both cases, a sergeant seeks to "break you down in order to build you back up"— in a different way than you were before. This is preparation for war, in which you will be asked to do things normal civilians typically would be emotionally ill-equipped to perform, such as killing other human beings. When the soldier returns, now a war veteran, to his original civilian environment after the war-chaotic jungles or wastelands, he has a difficult time readjusting psychologically. Like the stepping-over robot, or mother who has deliberately cultivated herself to be ready to sacrifice her child at moment's notice, the returned veteran soldier has developed a different dispositional set: one that makes it difficult to maximize utility for himself or others in the original context once he's been cultivated differently in preparation for a foreign environment or circumstance.

[360] This example was offered to me by my dissertation advisor, Norman Dahl, and nicely elucidates the distinction between action and motivations.

right. Consider an example where a doctor gives medicine to a child to save his life, but only because the mother is his ex-wife, whom he viciously hates. The mother of the terminally ill child despises and resents the child, and is gleefully waiting for the child to die in order to collect upon the child's life-insurance. The hatred of the child is actually the only thing the doctor and his ex-wife agree upon. But the doctor is willing to overlook his own hatred for his child in order to spitefully harm his ex-wife. This is an action that is morally right, by act utilitarianism, but is still an action that we would view as blameworthy – we would blame the doctor for not acting on different motives, and we would certainly not praise him.[361] On the other side – a wrong action that is praiseworthy – we can look at the hippo-god example, detailed above.

Notice in our discussion above how it is difficult to distinguish action from motive. The spiteful doctor's action, while the morally right action, by act utilitarianism, is blameworthy. But it's not really his action in isolation that is blameworthy: more to the point, it is his disposition that is blameworthy, which makes his action, though utility-maximizing in this instance, a blameworthy one. Contrarily, the mother's action in the hippo-god case, though morally wrong, is a praiseworthy action (presuming we can conceive of it as coming from a person who is functioning well, morally, given their background).

In conclusion, we should be sure to maintain this distinction (though it may seem obvious in retrospect) that a person can be praiseworthy or blameworthy in their disposition, but that a person is neither praiseworthy nor blameworthy in their actions until they actually act. As soon as the person acts, his action can be assessed as morally praiseworthy or blameworthy, though not in terms of whether the action is morally right or wrong itself, but in terms of whether the agential action is a natural expression of a morally praiseworthy or blameworthy disposition.[362]

---

[361] This action is *not* right by dispositional utilitarianism, as the spiteful doctor is morally unfit: his disposition is one that is not likely to maximize utility over a lifetime. He is thus a morally bad person. Had the action been in maintenance or cultivation of a utility-maximizing dispositional set, it would have turned out to be the right action.

[362] Related to this, a person can act contrary to their character: such as a generally benevolent person insulting a friend who's gotten on his last nerve. This action is in fact blameworthy, though it is blameworthy as an extension of the part of their disposition that allows such a lapse in benevolence.

Blameworthiness: Actions versus Dispositions

Blameworthiness is often attributed to an agent's actions, not just their dispositions. Consider the example of a married man who ends giving in to the temptation of having an affair with a coworker, thereby betraying his wife. This man, call him Bill, might be considered a very thoughtful and measured person, who almost always refrains from violating his morals, commitments, and integrity in the face of temptation. If Bill, save this singular departure from moral rectitude, is generally a utility-maximizing agent, then dispositional utilitarianism would even endorse him as being an altogether good agent. To clarify the issue of action-blameworthiness, consider a contrasting case: imagine William, who is identical to Bill, is a married man but avoids the moral violation of an extramarital affair simply because there is no particularly alluring coworker in his workplace, which would lead him to stray.

According to dispositional utilitarianism, both Bill and William would be equally blameworthy in their moral character, given their blameworthy sub-optimal dispositions. Nevertheless, in another way, it appears that Bill is blameworthy whereas William is not: Bill is blameworthy for his *action*, for cheating on his wife, whereas William engaged in no moral violation and thereby is not blameworthy vis-à-vis action. Does this not show that dispositional utilitarianism leaves something out of the moral equation, namely the blameworthiness of actions?

My response to this objection may seem unsatisfactory to our commonplace morality: I deny that actions, by themselves, are praiseworthy or blameworthy. I would maintain that Bill and William are equally blameworthy as far as character, and that their respective actions are only relevant as being indicative of that character. Bill just happens to be in an environment where his sub-optimal dispositional set manifests in his betrayal of his marriage vows. William, on the other hand, is in an environment where there is no such catalyst for this betrayal. Both would betray, however, given the same circumstance. While Bill could be colloquially described as "acting out of character," all this seems to mean is that his disposition is such that in, say, 99 percent of cases where temptations of

infidelity are present, he would resist such temptation. The fact that Bill is in a situation where this one percent obtains does not mean he's acting out of character, but that this is the one circumstance out of a hundred where we can see he does not have an entirely perfect character in regard to his fidelity. Still, the point remains that Bill intuitively seems more morally blameworthy than William: after all, Bill committed a moral offense whereas William did not.

I concede that Bill should be held as blameworthy by society, his wife, his peers, and so forth, whereas William should not (having committed no moral offense). Perhaps, however, there are both an epistemic and pragmatic reasons for blaming Bill more than William. Epistemically, outside of our hypothetical examples, we cannot actually know what a person's character is without surveying their actions over the course of time: actions indicate character. Pragmatically, perhaps we should react negatively to those who commit moral offenses, and pretty much leave alone those who do not commit such offenses; we are probably better off approaching blameworthiness in this way. Nonetheless, I would maintain that it is not one's actions, by themselves, that makes one blameworthy: it is the dispositions that lead to such actions.

I recognize that this position represents a counterintuitive point of view regarding praiseworthiness. In defense of the plausibility of my view, I would appeal to the reasoning Thomas Nagel (1979) presents. Nagel demonstrates that our intuitions regarding blameworthiness are inconsistent, given that character dispositions of the agents in each case are the same. Take, for example, the case of two drivers, Al and Ben, who are separately driving down identical streets, heeding all legal requirements, except that they are perhaps driving 3 miles per hour over the speed limit. As Al is driving, a child suddenly runs into the middle of the street, trying to retrieve a ball. Al swerves, slams on the brakes, and does everything to avoid hitting the child; unfortunately, he hits the child with the car, thereby killing the child. Ben is identical to Al, and so are the circumstances, save one happenstance: there is no child that runs into the middle of the street. As a result, Ben kills no one. In evaluating these cases, we tend to judge Al as blameworthy more so than Ben. Yet this seems an odd disparity: the external

circumstance of the child was out of Al's control, and it was just "bad luck" that a child ran out into the middle of the road in his case.

Another case we might consider is where two identical mothers, in two different possible worlds, say, are giving baths to their 3-year-old and 5-year-old. Needing to quickly put some washed clothes into the dryer, both mothers momentarily leave their two children in the bath unattended for three minutes. When the first mother returns, the 3-year-old and 5-year-old smile up at her, happily splashing around, while the second mother returns to discover in horror that the 3-year-old has drowned, and the 5-year-old is trying to drag him out of the tub. This scenario presents a case where the agent makes a voluntary choice to leave young children unattended in the bath, yet it seems we would more readily and severely blame the second mother over the first.[363] I would venture that many "good" parents – generally responsible and loving – have, at times, left young children unattended in the bath or in other slightly dangerous situations, yet we would presumably not hold all of them as blameworthy as the mother whose action actually led to the death of her child.

In consideration of this phenomenon of moral luck, Nagel presents four types of moral luck: resultant, circumstantial, constitutive, and casual. Resultant moral luck is represented in the scenarios above. I will briefly visit the three remaining types of moral luck, though it is not my intent to fully explore the subject; I merely intend to show how a possible defense can be mounted in support of the position that blameworthiness is appropriately attributable to dispositions but not actions by themselves.

Circumstantial moral luck is a second kind of moral luck that relates to the circumstances one is in: for instance, Germans who persecuted Jews during WWII or knowingly allowed such persecution to occur, are morally blameworthy; however, if they had moved away before the onset of this persecution, they would not be in the circumstance where they would have been involved – either actively or passively –and thereby they would not be morally blameworthy. However, these circumstances are mere moral luck. In fact, psychologist Stanley Milgram demonstrated through his famous

[363] Especially when these scenarios are presented singularly, rather than together, where a person might note the disparity in blameworthiness, and seek to bridge this gap by attributing more blame on the lucky mother.

"learning" experiment, that most people, Americans this time rather than Germans, seem capable of being as callous and cruel as the Nazis; yet, the experiment volunteers do not seem as blameworthy to us.[364]

Constitutive moral luck regards a person's character, citing that an individual's moral character is largely determined by forces outside their control. And yet we oftentimes attribute blameworthiness to an agent, irrespective of these forces. It seems that in some cases that these outside forces should be taken into account when assessing blameworthiness.[365]

Causal moral luck relates to determinism: Is blameworthiness a morally sensible notion if a person who acts badly could not have done otherwise? I will not pursue this line of thought much here, as this form of moral luck is often viewed as redundant, provided the other three forms of moral luck that Nagel develops, and I also wish to avoid delving too deep into the complications of the free-will debate.

The phenomena of moral luck, as characterized above, presents a dilemma: either individuals are morally responsible for all of the actions they do, even involuntary ones, or they are morally responsible for none. There have been various pragmatic responses to this dilemma. One is proffered by Susan Wolf (2004). Wolf incorporates a "rationalist" and "irrationalist" approach, in attempt to reconcile the dilemma. The rationalist approach parallels the dispositional utilitarian position: equal dispositions deserve equal moral blame/praise, irrespective of what consequences happen to manifest due to moral luck. The irrationalist approach holds the equal dispositions do not deserve equal blame: agents of unlucky consequences deserve more blame than identical agents of lucky ones. She argues that the unlucky agent ought to accept that he has a special connection to the consequences, as he is causally responsible for them. Because of this, she argues, the

[364] I would speculate that if the electro-shock was real, we would probably consider the majority of the volunteers who inflicted lethal voltage to the "learner" as morally reprehensible. Yet interestingly we tend to give no thought toward those who would have possibly killed the learner, in part, because the electro-shocks just so happened to be fake.

[365] For instance, we might imagine a man who has "road rage" issues due to his brain chemistry and/or upbringing, and that despite his voluntarily and diligently attending anger-management classes in an effort to get in control of his impulses, he ends up beating up another motorist who cut him off. Another driver might be completely serene in his driving, given his brain-chemistry and/or upbringing, and so does not commit such an offense.

unlucky agent must hold himself as more morally blameworthy than the lucky agent would be. Wolf holds, nevertheless, that outside evaluators should hold them equally blameworthy, despite our intuitions to the contrary.[366]

My intention is not to necessarily endorse Wolf's argument. I merely wish to illustrate that there are some attempts to reconcile the dilemma, which maintain that the blameworthiness of others should be taken solely in regard to their dispositions, however counterintuitive this may be, while maintaining that there may still be a way to attribute more blameworthiness to agents for actions that happen to result in unlucky consequences.

Dispositional Utilitarianism and Praise and Blame

According to dispositional utilitarianism's definition of the "right action," the right action and the praiseworthy action are nearly always correlated – in a similar way to how fitness and function are highly correlated in evolutionary biology. "Right action" is defined as that action which maintains or cultivates a dispositional set that could be reasonably expected to maximize utility over the course of the person's life when compared with acting from some other dispositional set. This high correlation between right action and praiseworthiness seems to be a virtue of dispositional utilitarianism, as it will tend to avoid counterintuitive instances that act utilitarianism must countenance: where right action and praiseworthiness conflict more commonly. Dispositional utilitarianism must still countenance the fact, nonetheless, that an individual can sometimes perform the wrong action, be a bad agent, and yet still be praiseworthy,

[366] Again, I think there might be good pragmatic and epistemic reasons for holding people as more morally blameworthy for bad moral luck: for example, there might be good pragmatic reasons to sentencing a criminal who accidentally shoots an innocent bystander to more prison time if that bystander dies in the hospital as opposed to survives after surgery. There might be good epistemic reasons for this as well: the death of the bystander seems an indication that the criminal is malevolently negligent or even hostile toward the innocent; we honestly do not know to what degree the death of the bystander is just anomalous bad moral luck, or par for the course given the criminal's disposition.

An agent could be said to be praise/blameworthy if she has a utility-optimal dispositional set that has been cultivated via her development background. For instance, a suburban "soccer-mom," is functioning well, morally – and is thereby praiseworthy (or at least not blameworthy) –if she has developed in her cultural context a dispositional set that maximizes utility: say, for instance, if she makes choices and sacrifices that benefit her child insofar as he can flourish, but doesn't overdo it at the expense of others.

In contrast, consider the example of an Inuit mother who – when necessary to ensure tribal survival, and with reluctance and grief – voluntarily leaves her child to die in the snow of hypothermia. If the background conditions have been such that the mother needs to make such a sacrifice, in order not to imperil the rest of the Inuit tribe, then we should say that that mother both functions well, morally, against that developmental background, and has a utility-productive dispositional set. The Inuit mother functions well in that she has a productive dispositional set that has arisen from that culture due, at least partly, to recognition of consequences.[367]

The degree of praise/blameworthiness of the Inuit mother is in proportion to her functioning well, given her developmental background.[368] The Inuit mother would not be blameworthy if she were taken from her society and put into a novel environment – say, an upscale American suburb – even if her dispositional set was far below utility-optimal in that new environment.[369] For instance, the Inuit mother might have a steeled, if unhappy, readiness to sacrifice her child for greater welfare of her neighbors.[370] Her

---

[367] A morally well-functioning individual will, presumably, need not have a utility-optimizing dispositional set; presumably, the individual, for instance, could do more to diminish world hunger than she is doing. A culture might say she was functioning *better* if she did these "supererogatory" acts. The point here is that dispositional sets and moral functioning are related to each other: the first being utilitarian in reference, the second being more general and non-theoretical in reference.

[368] A person would also be praiseworthy if she were acting in a way that a morally well-functioning person would be acting, though she might not be morally well-functioning herself overall. This is relatable to a person who performs that action that a virtuous person would do – for the right reasons, at the right times, with the right feeling, and so forth – though she may not be morally well-functioning herself overall.

[369] In brief analogy to evolutionary biology, an animal put in a novel environment could be well-functioning, yet unfit. For instance, if a polar bear is placed in a forest, its fitness decreases since its white fur no longer affords it a predatory advantage vis-à-vis camouflage. Nevertheless, the polar bear is still, presumably, well-functioning.

[370] I assume we'd have to do some tinkering to make this example quite right: for instance, that the Inuit mother thinks of her neighbors in the same way as she does her tribe, without bringing in a different background.

readiness is not blameworthy, however, given its historical background: she is functioning well, though it is not a comparably good dispositional set for the new environment.

In contrast, imagine if an American soccer-mom swapped places with the Inuit mother: the soccer-mom would not be blameworthy, by commonsense morality, in refusing to abandon her infant to the ravages of hypothermia; rather, I would think we would find her blameworthy if she could readily abandon her baby. Our moral condemnation of her, were she readily able to abandon her baby, would seem unmoved by the fact that her sacrifice protected the precarious and dire welfare of the tribe of which she was now a member. The soccer-mom, thrust into this new context, who was emotionally able readily to sacrifice her baby, may have a utility-optimizing dispositional set, but would be morally functioning poorly given her background. The soccer-mom might lucidly understand the necessity of sacrificing her baby for the good of the tribe, and perhaps could somehow manage to do so by overcoming her emotional attachment to her child: such a story might make her not blameworthy or might, arguably, make her even praiseworthy by making such a sacrifice while confronted by the cold hard facts. Nevertheless, her being "readily" emotionally able to do so would show her morally functioning poorly, given her background.[371]

Likewise, in a case where a mother cannot bring herself to boil her baby. Despite the fact that boiling her baby would lead to a better state of the world, we could not say the mother would be functioning well if she could make such a morbid sacrifice at the altar of utility. We would evaluate the mother as praiseworthy if she were unable (or at least not blameworthy, certainly), as the mother would be functioning well, by moral lights. This may seem a strange set of assertions: The mother is praiseworthy but unable to perform the utility-maximizing action. But remember that if she were able to perform the action, it would not be the "right" action: Again, a right action, as defined by dispositional utilitarianism is an action is right if and only if it maintains or cultivates a dispositional set that could be reasonably expected to maximize utility over the course of

[371] The soccer-mom might be able to bring herself to make this sacrifice over time, in a similar way that Huck was able to bring himself to refuse to betray Jim, despite his own conscience. To be able to do so immediately, however, would be an indication of not functioning morally well.

the person's life (in that environment) when compared with acting from some other dispositional set.

An agent, however, does not need to consciously recognize, and be deferential to dispositional utilitarianism, or any utilitarianism; in fact, it might in fact be best that she not be.[372] Instead, it's probably for the best that she be disposed to follow certain moral precepts with escape clauses that activate at certain thresholds or certain contexts: For instance, she might not be willing to push a man off a footbridge to save five people, but might be so willing if it were to save a busload of a hundred schoolchildren. The degree of this threshold will depend upon context: for instance, the soccer-mom should have developed a much higher threshold than the Inuit mother when it regards sacrificing her baby, given a difference in her environment.

While commonsense morality recognizes praiseworthiness, dispositional utilitarianism, and most utilitarian theories, might be thought of as blind to moral function and praiseworthiness. After all, dispositional utilitarianism seems only to be concerned about the utility produced by dispositional sets, just as act utilitarianism, more traditionally, is only concerned about the utility produced by agential action. One common criticism of act utilitarianism, in fact, is that it is unjust when it comes to punishment: after all, act utilitarianism would just as readily punish the innocent as the guilty if the utility calculus worked out right. It is unconcerned about desert; it only concerns itself with future consequences and their effect on utility. Because of this, while act utilitarianism is able to deem the right action an agent can take, it seems to be blind in regard to praiseworthiness and blameworthiness of that agent. Praiseworthiness and blameworthiness is sometimes thought not to fit into the act utilitarian paradigm, as these terms appear to regard kind of accolades or disapprobation a person *deserves*.

As should be evident in our discussion so far in this section, I do believe utilitarianism has a place for praise/blameworthiness. Dispositional utilitarianism can

[372] The agent should be acting deferentially of course, at some level, to morally relevant factors – such as consequences, harms and benefits, rational rules of actions (e.g., the Golden Rule). She just needn't be a utilitarian. Similarly, virtue theory doesn't necessitate that an agent act out of conscious recognition, while performing an action, of what character traits will likely enable a flourishing life and make one good qua human beings: however, the agent's action should be an expression of a certain virtuous dispositions that defer to these factors.

hold someone as praiseworthy if they have the productive dispositional set that could be reasonably expected to have given their developmental background. Happily this dovetails with what is meant by "moral function" as previously described. A person is blameworthy, by dispositional utilitarianism, if they do not have a productive dispositional set that they could be reasonably expected to have given their developmental background. The suburban soccer-mom would could readily sacrifice her baby to the runaway trolley-car has a dispositional set that would ordinarily be nonproductive, given her developmental background. The Inuit mother who could readily sacrifice her baby for the significantly greater welfare of her tribe has a dispositional set that would ordinarily, in fact, be productive given her developmental background, and for this reason would not be blameworthy (and might even be praiseworthy).

We might consider such rare cases, where an "agent" might fail to do the right action, defined by dispositional utilitarianism, and yet still seems praiseworthy according to both dispositional utilitarianism and commonsense morality. A "right" action for a person, according to dispositional utilitarianism, is that action that is maintains or cultivates a dispositional set that could be reasonably expected to maximize utility over the course of the person's life when compared with acting from some other dispositional set, in that environment.. The soccer-mom, suddenly displaced into Inuit culture, for instance, would now be in an environment where the optimal dispositional set of a mother would be to be willing, though reluctantly, to sacrifice her baby for the precarious welfare of her tribe: the right action, then, in this environment would be to sacrifice her baby.[373] Though dispositional utilitarianism would deem this sacrifice as the right action, it still would not hold the soccer-mom to be blameworthy in lacking this optimal dispositional set in this new environment. Dispositional utilitarianism may prescribe that she change her dispositional set, through struggle and time, to a more utility-increasing set, presuming such an emotional and psychological change were possible. Dispositional utilitarianism, nevertheless, would not blame her for not having the optimal dispositional

[373] We would have to imagine that a "callous mother" toward her child might have a better dispositional set in the Inuit tribe's circumstance than the well-functioning mother. In such a case, dispositional utilitarianism would endorse the callous mother's set in this new circumstance (though not in the previous circumstance, where the dispositional set was cultivated).

set. Dispositional utilitarianism and commonsense morality would agree, then, that the displaced soccer-mom should not be blamed for her inability suddenly to make such a morbid sacrifice of her child, irrespective of the dire consequences possible without the sacrifice.

To return to the previous mother-baby case, act utilitarianism might be correctly presumed to prescribe a mother boiling her baby, if the utility produced via the state of the world resulting from that macabre action outweighs her not doing it (including her becoming an agent with a less-utility-producing agent over her lifetime). In order for dispositional utilitarianism to prescribe this action, however, it would first have to endorse this dispositional set, generally, given the background context of the world in which the mother lives. Dispositional utilitarianism cannot prescribe an action without endorsing an accompanying dispositional set in which such an action is possible. In this way, utilitarianism will endorse dispositional sets, based on the probable manifestation of actions that will lead to optimized utility given the particular world in which the agent will be acting. In a world, for instance, where such scenarios as the footbridge trolley scenario commonly occur, dispositional utilitarianism would likely endorse dispositional sets where an agent is readily willing to push the man to save the five. In a world where the footbridge trolley scenario very rarely if ever occurs, like our world, it would likely not endorse dispositional sets where agents were so ready, as presumably this readiness precludes other utility-productive dispositions.[374]

It's crucial to notice that counterexamples against utilitarianism often proceed by typically dropping bizarre scenarios into the middle of an ordinary world. It is then presumed that utilitarianism would endorse taking the putatively immoral action, such as pushing the man off the footbridge, or framing the innocent, and so forth – especially with the improbable addendum that it's guaranteed no one will ever find out.[375] In initial response to such examples, it should be noted that utilitarianism need not endorse this putatively immoral action if it means the agent would not have a productive dispositional

[374] In addition to social inculcation, if such a world was the case, we might speculate we would be different evolved beings, with possibly different kinds of moral intuitions (e.g., more directly utilitarian).

[375] I will consider the "unrealism" of such examples further, in reference to our fourth example: framing the innocent man.

set, in relation to utility-optimization given an ordinary background. I proffer that the intuition against pushing the man off the footbridge resonates in recognition that the agent, if he did so, would not be morally functioning well, given this ordinary context, and that this matches up in relation to dispositional sets.[376]

But if we consider, in contrast, another context, where readiness to make such sacrifices is to function well, the counterintuition doesn't seem to arise, or at least not as quickly or assuredly. For instance, in a time of war, morality doesn't seem to apply in the same way, and our intuitions tend to be different. Murder is a word that is less commonly used in a time of war than a time of peace, despite warfare where recognized innocent civilians are collateral damage. In a time of war, moral acceptability seems to shift in the next context. Consider the example of how medical doctors in a MASH unit might adopt a policy of "triage," and might be more likely to sacrifice one severely wounded patient for the better good of the other five.[377] The same doctors during peacetime would not practice according to triage principles.[378] There are cases, during the Vietnam War, where a soldier might shoot a child running toward his platoon due to justified suspicion that the child was carrying a grenade. Other cases of armed conflicts involve enemy combatants using a child as a human shield, where a soldier would shoot the child, in order to prevent the enemy combatant from killing a number of his fellow soldiers. Presumably, this sounds reluctantly acceptable to us, morally. In contrast, in a non-wartime case, such as contemporary America, it seems we would not find it morally acceptable to make this trade of human lives – for instance, a hostage situation where a

[376] Again, dispositional utilitarianism doesn't care about moral function; it only recognizes dispositional sets. Typically, dispositional sets and moral function are unified. To reemphasize, I am not asserting function as morally correct; I am merely asserting it is the underlying concept for praise/blameworthiness. In this way, utilitarianism doesn't have anything necessarily to apologize for in not recognizing moral function. An argument could be made, of course, that moral function is the correct moral concept, which utilitarianism fails to recognize, but I will not engage that possibility here. I am asserting that dispositional sets, given the developmental background remains constant, is just as good as moral function. To represent it by a simple equation: dispositional sets + developmental background = moral function.

[377] Peter Singer explores this wartime policy in analogy to feeding starving peoples in developing nations, in his article, "Rich and Poor," (1993b, pp. 235-239).

[378] This wartime example parallels a common act utilitarian counterexample in which a surgeon, presumably in an ordinary context, must decide whether or not to harvest the organs from one savable patient in order to save the lives of five others.

bank robber was using a child hostage as a shield – despite its allowing him to kill a number of other civilians. In the wartime case, it seems the soldier can be functioning well, and thereby not blameworthy, whereas in the peacetime case, it seems a police officer would not be functioning well to kill the child to save other civilians.

Motive-Blindness: An Objection to Utilitarianism

As previously discussed, one problem that utilitarianism faces is its apparent blindness to motives. This is especially a problem for act utilitarianism: The spiteful doctor could save a child's life while hating the child, because he hates his ex-wife more so. By act utilitarianism, this doctor would seem to be just as laudable as a sympathetic doctor who saves the child's life for the good of the child: both actions lead to identical consequences. I believe that dispositional utilitarianism, although suffering some degree of motive-blindness itself, has more resources than act utilitarianism to surmount this difficulty.

In the case of the spiteful doctor, dispositional utilitarianism would ask if the doctor is morally fit: Does he have the disposition likely to maximize utility over his lifetime? It seems unlikely a person who is single-mindedly spiteful and callous would be so disposed toward utility maximization, whereas the sympathetic doctor presumably would be far more disposed.

We might imagine another example on this point: an ostensible altruist who performs a lifetime of good deeds motivated solely by an egotistical feeling of superiority in comparison to everyone else, where he doesn't even truly care a whit for the destitute whom he is helping (though he would never admit this).[379] We can ask of the egotistical philanthropist the same question as the spiteful doctor: Would he have a disposition likely to maximize utility over a lifetime?

[379] I suspect the reason we find certain motivations to be praiseworthy is their tendency to lead to benefits in a large number of cases. The man who dedicates his life to helping others solely so that he can feel superior to others who are less charitable does not seem realistic to me, however. Even if it is a realistic characterization, we imagine that this person would not be maximizing benefits in his life, given his ultimate self-centeredness, lack of true sympathy to others, and his need to feel superior to his peers.

The objector to utilitarianism might be hard-pressed to generate intuitively palatable cases. The cases above don't seem satisfactory: we would have to believe that the spiteful doctor and egotistical philanthropist were both equally disposed as their sympathetic twins. It seems unlikely that there will arise many circumstances where an agent, only acting out of spite and with an intention to emotionally harm, will actually maximize utility. It similarly seems unlikely that a person who is motivated by a desire to feel superior, with no mind toward those he's helping and even no sympathy toward them, will actually maximize utility over a lifetime in a typical environment. It seems very unlikely utility will be maximized for himself or for others. It's possible, of course, that we can cook up a scenario where the spiteful doctor and selfish philanthropist *do* happen to maximize utility over their lifetime, but moral fitness relates to the *probability* one's disposition will maximize utility over their lifetime, not how much utility is actually maximized over their lifetime.

In order for the spiteful doctor and the egotistical philanthropist to serve as legitimate counterexamples, they must actually be plausible, and not just logically possible in our imaginations. In the very least, the burden of proof seems to be shifted onto the objector to generate a plausible scenario where, generally, an agent has morally bad traits, in our commonsense moral evaluation, and yet is still endorsed by dispositional utilitarianism. The difficulty of meeting this burden in the case of dispositional utilitarianism, compared to that of act utilitarianism, shows its advantages in explanatory power. After all, act utilitarianism would say that the sympathetic philanthropist giving to charity is the same, morally, as the egotistical philanthropist giving to charity. Act utilitarianism has no resources to make this distinction: they both equally maximize utility. Dispositional utilitarianism has further resources to differentiate the two because, presumably, over a greater span of time, their characteristics will manifest in disparate actions: sympathy leading to greater utility, and egoism likely leading to lesser utility (why else would we devalue self-absorption and callousness if it

did not lead to harms to of some sort?). In this way, dispositional utilitarianism can better stave off this objection from motive-blindness. .[380]

There is another possible response to the charge of motive-blindness: namely, that dispositional utilitarianism need not be blind. One possible way of removing this blindness is to add an addendum that one's actions should not only proceed from a dispositional set that would maximize utility over a lifetime, but that, further, the individual must recognize why she ought to act from such a dispositional set. I am somewhat resistant to this addendum, however, in that it seems – at least at first – to necessitate that a person subscribe to utilitarianism: that they recognize that the reason she ought to act from these motives is toward the end of maximizing utility over her lifetime. I believe the addendum can provide a helping hand in a more general way, however, and that the moral agent need not be so strapped to utilitarianism, specifically.

We can start from a fundamental difference that all moral theories must make: the difference between a moral agent and a brute cause. A moral agent has intentions of some sort: a runaway trolley does not. The first is praise/blameworthy: the second is not. It seems that in order for a theory to be an adequate moral theory, it needs to make this sort of distinction between brute causes and (moral) agents, with intentions and voluntary choices. Part of agency involves reflection: the moral agent moves beyond being a mere agent in her recognition of and reflection upon the moral dimension of her actions. She then becomes a moral agent. Presumably most animals do not do this.

In defense of utilitarianism against motive-blindness, then, I would say that utilitarianism must recognize that any moral agent must be acting – not necessarily from utilitarianism motivations– but from *some* moral motivations: at least in reflection of moral motivations rather than in obliviousness to them. This leaves the range of

[380] We might consider an extreme example motive-blindness, which I will call "the inverted moral spectrum" case: Imagine that, due to faulty cognitive wiring, an agent has the misunderstanding that human pain behavior denotes pleasure, and human pleasure behavior (e.g., smiling, laughing) denotes pain. This agent seeks to cause optimal pain and suffering in the world but, due to his inverted cognition regarding language, human behavior, etc., the agent ends up mistakenly maximizing utility. His motives are horrific, but his dispositional set is utility-optimizing. Having such a utility-optimizing dispositional set, his set would be fully endorsed by dispositional utilitarianism. This example seems farfetched, for we would have to imagine an agent (perhaps an alien?) making fundamental and deep mistakes about human behavior: believing feeding the starving causes pain, that rescuing the drowning causes pain, and etc.

motivations for moral actions quite broad, the only limit being that a moral agent – to be a moral agent – must be motivated by what he believes to be morally relevant features. This excludes the skeptic, the nihilist, and the evil-doer (the first two are in denial of these features and the third is in defiance of them). This minimal requirement for moral agency – that a moral agent have moral reasons – removes a portion of this blindness which would trouble dispositional utilitarianism, even if (it seems to me) only in implausible cases (such as evil or accidental agents that happen to be disposed to maximize utility).

In any event, it is not my intention to provide an exhaustive defense of utilitarianism, and so I will not further explore this objection. For my purposes, it should suffice to say that dispositional sets among human beings will typically involve motivations that we would recognize as praiseworthy.

I will discuss the notion of dispositional sets further in the next section, while shifting toward the third caveat, boundedness, in reference to another famous thought-experiment presented by Williams in objection to utilitarianism.

Williams vs. Hare: The Plane Crash Case

In a televised debate between Professor Bernard Williams and R. M. Hare, Williams challenged Hare with the following thought-experiment:

> "You are in an air crash and the aircraft catches fire, but you have managed to get out; in the burning plane are, among others, your son and a distinguished surgeon who could, if rescued, save many injured passengers' lives, to say nothing of those whose lives he would save in his subsequent career. You have time to rescue only one person" (1981, p. 138).

In a similar way to the trolley case, above, the plane crash example presents utilitarianism with a question: Who should the father save, his child or the surgeon? Much like the trolley example, this is a numbers game, yet the question is who to save, rather killing versus saving. This removes the doing and allowing distinction from the ethical equation.

Should the father save his child or save the surgeon? Williams asserts that utilitarianism would recommend the father save the surgeon rather than his child, thereby enabling the surgeon to save the lives of more passengers. Is William right in his assertion of what utilitarianism would prescribe? Given the broader scope of dispositional sets – which manifest in a series of actions over a lifetime – it seems at least arguable that utilitarianism would prefer the father with a dispositional set that makes him incapable of not rescuing his own child from imminent death in order to save the life of the surgeon.[381]

If our moral judgment is sensitive to agential function – which I believe our moral judgments are – then counterexamples leveraging isolated utility-maximizing actions against what is commonly considered good moral character will be diminished in their impugning force: namely, because in many cases utilitarianism will endorse the agent functioning well, in terms of utility-maximization, as it tends to lead to greater utility over an agent's lifetime. For instance, in the plane-crash case, utilitarianism would

[381] A question that will naturally arise is at what point the threshold lies where such a paternal sacrifice would not indicate a morally degraded dispositional set. For instance, we might imagine that a father was faced with a decision to sacrifice his child in order to save a million children across the world; at some point, it seems, numbers do matter, and an agent with a morally upright dispositional set will make such a sacrifice.

arguably endorse that the father be the sort of father who saves his child over the surgeon. If the father saves the surgeon over his child, then he lacks a morally upright disposition – one which is also not utility-maximizing, in ordinary contexts throughout a life, in comparison to the disposition of the father who would save his child instead.

Such counterexamples which play upon singular instances, claiming utilitarianism need endorse the utility-maximizing option in the short-term, are myopic – they fail to take into account all of the morally relevant features to which our moral intuitions are sensitive: namely, dispositional sets (as well as proper moral function). According to dispositional utilitarianism, an individual should perform that action that would by done by a person with a dispositional set that would likely maximize utility over a lifetime given the environment. The "right action" in this case then, according to dispositional utilitarianism, would be for the father to save his son rather than the surgeon. Dispositional utilitarianism, then, sidesteps Williams' criticism by not endorsing the counterintuitive prescription. Furthermore, neither dispositional utilitarianism nor commonsense morality would blame the father for "failing" to perform the action that would maximize utility. This is closely similar to the suburban soccer-mom who is unable to hurl her child under the wheels of the runaway trolley-car in order to save five lives. Now considering the displaced soccer mom who is readily able to sacrifice her baby for the greater good of her precariously-situated Inuit tribe is morally functioning poorly, but making this sacrifice would still be the right action, given that it would be the action done by an agent who had a dispositional set likely to maximize utility in that environment. Nevertheless, she would not be praiseworthy for this right action. In fact, she would, arguably, be blameworthy given that she is functionally poorly, morally, given her developmental background: our moral intuitions would blame (or at least not praise) the displaced soccer mom for her readiness, though utilitarianism would endorse her dispositional set in this new environment, as well as her sacrificing her child in this new context.

This example might be thought to serve as a compelling counterexample against dispositional utilitarianism, given that it pits right action versus praiseworthiness, where the two are at odds. I would disagree for three reasons. First, this alleged counterexample

seems paltry, as this situation, involving a significant displacement of the agent from her developmental background, seems so rare a case that we might reasonably doubt that we can rely upon the deliverances of our moral judgment. After all, our moral cognitive faculties, presuming they are truth-tracking, have been developed culturally and evolutionarily to deal with certain contexts.[382] If we take them out of these contexts, we should be less sure of their fidelity. Second, we seem to recognize the trickiness of moral evaluations in other cultures that are different from ours, for instance, where some cultural practice (such as facial scarring rituals) might be deemed morally good or permissible in that other culture but not in our own (and vice versa). This recognized difference between environments does not lead to moral relativism, though it does necessitate deference to "context-sensitivity," where differences in the environment may understandably necessitate corresponding differences in moral dispositions and moral practices. If an agent transversed cultures – traveling to our culture from his – and practiced facial scarring on an American child, we would think him to be performing the wrong action, given the environment, but might deem his disposition not to be blameworthy (or at least less so) if he was trying to benefit this child. This seems to in fact square with our commonsense moral intuitions rather than run contrary to them. Third, I believe this conflict is also present in virtue ethics, which also allows for some degree of moral relativity: A man from a hunter-gatherer society might be virtuous in his own culture while being excessively bellicose in our own. In our moral assessments, then, it is necessary to be mindful of environmental contexts, developmental backgrounds, and the associated limits of our moral evaluation. This introduces the third caveat, which I am referring to as "boundedness" which I will explore this concept further in the next section.

So far we've seen two ways objections based upon counterintuitions can fail. First, the supporting counterintuitions may be noncredible if they are not based upon morally relevant features; consequently, these counterintuitions may be dismissed as

[382] A compelling analogy for this might be our intuitions predicting the behavior of physical bodies in space. In gravity-laden situations, such as on Earth, we have generally reliable intuitions about how physical bodies will behave; however, in gravity-free situations, our intuitions are far less reliable. Likewise with physics in a vacuum: consider, for instance, dropping a feather and brick from an equal height, and our initial intuitions regarding which will hit the ground first.

counterevidence to moral principles/theory. Second, even if counterintuitions are thought to be credible, the intuitions may be misinterpreted regarding what they actually refer to: for instance, intuitions may be sensitive to dispositional sets and not merely to singular actions. Misinterpreted intuitions cannot be used as counterevidence, just as misinterpreted data in science cannot be used as counterevidence against a hypothesis.[383]

The third caveat, "boundedness," is proffered by Hare (1981). He claims that moral intuitions are bounded in that they are socially (or evolutionarily) inculcated to deal with a limited range of cases. Anything outside that range, much like the constraints of an experiment,[384] may not result in reliable intuitions: that is, the intuitions generated may be flawed data which do not map the moral features of the world. In relation to the plane-crash scenario, Hare explains that, as the father:

> "You will almost certainly rescue your son. But that is because you have (rightly from the critical utilitarian point of view) been brought up to attach dominant importance to these family loyalties. Of course no upbringing takes into account such rare cases as this (they are not what those who influence you were preparing you for, nor would evolution be affected by them)" (1981, pp. 138-139).

Hare's response can be interpreted to suggest that, given the father's societal and evolutionary context, the father has been instilled with a certain dispositional set, including, therein, moral principles and morally-valenced emotions. This rare case where sacrificing his child maximizes utility (similar to a mother boiling her baby case) is not a case where we would expect an agent's dispositional set to be properly prepared. As such, it seems no failure of the father that he would not be disposed to make such a sacrifice, despite its maximization of utility in this instance. By the lights of dispositional

---

[383] In illustration, we might revisit the example of a credible data set of recorded movements of an astronomical body, and yet this data set might be misinterpreted as referring to the orbit of a planet rather than that of a moon, where the former interpretation would be at odds with a Newtonian hypothesis, whereas the latter would not.

[384] In science, for example, if experimental parameters are breached, then the resulting data may not be deemed credible, and subsequently cannot serve as a refutation or corroboration of the hypothesis. For example, the ideal gas law only applies to real gases at low pressure and high temperature, or for molecules without strong intermolecular forces. If an experimental setup violates these parameters, then the resulting data cannot be deemed credible: for instance, in utilizing the ideal gas law to derive the pressure of a gas chamber, if we knew the experimental parameters were breached, we'd have reason not to deem credible a mathematical derivation of the pressure of a chamber from looking at just the volume and temperature of the gas.

utilitarianism, to make this sacrifice of one's child in such cases is not in fact the "right action."

Hare's point segues to the third caveat of boundedness. In the examination of boundedness, I will consider both social/cultural and evolutionary inculcation – the latter of which Hare acknowledges but does not explore. In explication of boundedness, I will consider one of the most troublesome counterexamples to utilitarianism, which pits utilitarianism against our strongly held convictions of justice.

Framing the Innocent Man

One of the strongest counterexamples against utilitarianism concerns utilitarianism's supposed endorsement of injustice for the sake of optimized utility. An often-cited example in this genre is the "framing the innocent man" scenario from H. J. McCloskey (1965):

> "Suppose a utilitarian were visiting an area in which there was racial strife, and that, during his visit, a Negro rapes a white woman, and that race riots occur as a result of the crime, white mobs, with the connivance of the police, bashing and killing Negroes, etc. Suppose too that our utilitarian is in the area of the crime when it is committed such that his testimony would bring about the conviction of a particular Negro. If he knows that a quick arrest will stop the riots and lynchings, surely, as a utilitarian, he must conclude that he has a duty to bear false witness in order to bring about the punishment of an innocent person" (p. 127).

Despite the fact that framing the innocent man would prevent the suffering and death of hundreds of people, thereby maximizing utility, our intuitions tell us that framing the innocent man is morally wrong because it is unjust. This counterexample is one where all three caveats obtain: predication, dispositional sets, and boundedness.

Before fully engaging this example, the last caveat, boundedness, should be briefly elucidated. Boundedness asserts that moral judgments tend to be verisimilitudinous within certain parameters, yet decreasingly so as it departs these parameters. One infraction of boundedness I will call "displacement," where an intuition is being applied to a different context from which it has arisen, and to a context for which it seems ill-equipped. In analogy, consider our intuitions regarding the behavior of

physical objects. It seems counterintuitive to us, for example, that a feather and cannonball, simultaneously dropped at the same height within a vacuum will fall at the same rate. I would suggest that the reason we pretheoretically judge that the cannonball would fall faster is that our intuitions have been trained in an atmospheric environment, and our expectations presume those same conditions. Consulting our intuitions of physics when preparing for a mission to the moon, for this reason, would be imprudent; we should rather rely upon our principles – such as scientific theories – and not upon the intuitions that are not equipped for such foreign contexts. I will discuss displacement in a moral context later in this chapter when I revisit retributive intuitions.[385]

With boundedness in mind, we might begin our critical examination of the framing the innocent man counterexample in the way suggested by Hare: first by marshaling the utilitarian response regarding the unrealism of the example, citing that the hypothetical example is beyond the boundaries for which our intuitions are prepared.[386] Examples need to be within the normal conditions in which moral intuitions were originally inculcated. Hare emphasizes that an objector cannot use moral intuitions about

[385] Another infraction of boundedness might be called "multiplicative distortion." Our intuitions become less reliable if they are applied to quantities they are not equipped to assess. In analogy, consider a simple example where three jellybeans are on a desk. An adult looking at the desk need not count the jellybeans, from one to three, but can immediately apprehend that there are three jellybeans. This might remain the case from anywhere between one and ten jellybeans. However, ask a person to estimate the number of jellybeans a glass jar contains, and his intuition will likely be far less reliable when there are so many to behold, as opposed to when there are just a few. In moral philosophy, multiplicative distortion often arises in thought-experiments, usually in objection to act utilitarianism. In illustration, consider the following counterexample: Imagine a CEO and owner of a large cereal manufacturer were considering two possible ways to spend ten thousand dollars of his money (1) To spend the money to save five children dying in an impoverished community overseas: 5 children x 10,000 hedons = 50,000; (2) Or to use the ten thousand dollars to put an extra raisin in each box of Raisin Flakes: 1.2 million boxes x 0.05 hedons. Given the numbers, what should he do? Given that the utility of 60,000 due to raisins is bigger than the 50,000 due to dramatically improving the lives of five children, act utilitarianism would clearly prescribe that he add a single raisin to every box of Raisin Flakes, thereby maximizing utility. However, this seems the wrong answer: saving the five children just strikes us as morally better than adding a single raisin to a box of cereal, even if to a 1.2 million cereal boxes. At such points, it seems that math's reach has exceeded intuition's grasp. Can we really expect our moral intuitions to track such multiplication? I assert that we cannot depend on the fidelity of our intuitions to assess such cases of multiplication and large numbers. I suspect that multiplicative distortion may be at work to some extent in Scanlon's (1998) World Cup Match example.

[386] Hypothetical examples, it should be noted, are not necessarily unrealistic; and they are often a necessary tool to consider multiple possibilities before acting. Hypothetical examples, however, can quickly jump the fence, when their stipulations are too strong.

bizarre cases as artillery against utilitarianism, because in such cases our intuitions are significantly less reliable:

> "For his audience's intuitions are the product of their moral upbringings, and, however good these may have been, they were designed to prepare them to deal with moral situations which are likely to be encountered; there is no guarantee at all that they will be appropriate to unusual cases. Even in the unusual cases, no doubt, the usual moral feelings will be in evidence; but they provide no argument" (1981, p. 132).

Hare's comments relate to boundedness in the suggestion that our socially-inculcated intuitions, as well as evolutionarily-inculcated intuitions, are reliable when they function under normal conditions. These intuitions, however, are less reliable when they are displaced, and applied to novel contexts. Recollect our discussion of moral judgments condemning incest, retribution, and trust in chapter four.[387] We saw how novel contexts could diminish the credibility of moral judgments: these judgments were developed within particular contexts, and are likely to decrease in verisimilitude proportional to the degree of departure from the original context.

In analogy to science, scientific laws are expected to obtain only under certain conditions and within certain experimental parameters.[388] This example exceeds these boundaries in that it is a hypothetical example, where unrealistic factors are insisted upon. Life is not like this, however, and our intuitions didn't originate – socially or evolutionarily – under such conditions, and have not been inculcated to deal with such conditions. If the insistence of epistemic certainty were removed from this case, our intuitions and utilitarian's prescription would likely be compatible.[389]

---

[387] Regarding incest, we saw in the Mark and Julie case that no morally relevant features seemed present, though certainly moral judgments against incest are usually justified on MRFs. Concerning trust, we saw that politicians imitate behavior that typically stimulates the release of oxytocin, which increases evaluates of the person as trustworthy. I will consider moral judgments concerning retribution later in this chapter.

[388] For instance, the ideal gas law only obtains for ideal gases at high temperatures and low pressures, as the law does not consider the size of each molecule or the effects of intermolecular attraction.

[389] Of course, all thought-experiments in philosophy presume epistemic certainty where a helpful clause of "all things being equal" is presumed. In the Jim and the Indians case, for instance, we are assured that it is certain Pedro will honor his word, and spare the 19 other villagers. Even in scientific experiment, this *ceteris paribus* clause is inserted, where all possible interfering factors are held to be constant or accounted for, so that a particular phenomenon can be investigated. The problem is that epistemic certainty is a rare exception, and ideal conditions are hardly ever achievable. This practice of assuming "all things being equal" seems acceptable, however, presuming it is a reasonable assumption.

Our intuitions, presuming they are credible ones, originally arose because they were based upon morally relevant features.[390] In bizarre and extreme cases, however, we should expect that these MRFs may have changed, and we should be wary of this. The social and evolutionary inculcation of our intuitions were under particular conditions and within certain parameters.

This point was illustrated in chapter four, regarding our retribution intuitions. To briefly review: We might have the moral intuition that we should inflict righteous violence against criminals who egregiously violate social norms or harm their community. Evolutionarily, we are likely to have such normative intuitions because having them was in the interest of a community's functioning, and was thereby contributive to individual members' well-being, which ultimately tended to be in their genetic self-interest.[391]

Socially, we likely have this moral intuition because of cultural inculcation: it was in the best interest of individuals or their community if egregious norm violators were treated in this way.[392] Inflicting violence upon norm defectors deters them from defecting again, and deters other social cooperators from defecting initially. In a technologically developed society, the justice and penal systems replace the need for the interpersonal infliction of violence: institutions deter and punish by incarceration.

In brief illustration of this point, we might imagine an individual living in a hunter-gatherer society, where the infliction of violence (upon defectors who violate the

[390] For instance, if our intuitions weren't generally good guides regarding tendencies toward well-being, they likely wouldn't have been socially or evolutionarily inculcated.

[391] The evolutionary causal explanation for the presence of the normative intuition is in terms genetic-self interest. As explained in chapter 4, individual well-being tends to secure genetic self-interest. However, this is not always the case: for example, parents might have to make significant sacrifices of resources, negatively affecting their own well-being, in order to increase the chances of survival for their offspring.

[392] Social transmission of intuitions could occur in several ways; consider two: (1) social transmission of memes, where individuals in those groups who had learned particular customs/responses/intuitions gain selective advantage over individuals in other groups (2) conscious observation of how a custom/response/intuition is beneficial to the functioning of the group, and the decision to instill it in offspring. In example of the first, we can revisit the example of a culture's valuation that spiders are repulsive: we can imagine that individuals in groups that instilled repulsion to spiders would have selective advantage, *ceteris paribus,* over individuals in those groups that did not. In illustration of the second type of social transmission, we can imagine parents teaching their children to be afraid of spiders – in a similar way as "be wary/afraid of strangers" – and thereby conferring upon them selective advantage via social transmission.

norms of their community) is often to the end of social regulation. If this individual were to be placed in a technologically-advanced society which incorporates a justice and penal system, the infliction of violence upon defectors would no longer be to the end of social maintenance, but would be in fact disruptive and destructive. The morally relevant features between the two cultures differ: in the hunter-gatherer society, the intuition toward sanctions against defectors is based upon on morally relevant features that are absent in the technologically-advanced society: namely, the effect of sanctions to maintain community cohesion and engender the individual well-being of social cooperators.

The infliction of violence – via death penalty, whipping, hanging, or other methods – is no longer necessary in modern society due to a sophisticated justice and penal system. Nevertheless, our moral intuitions toward inflicting "justice" vestigially remain despite this novel context, much like phantom pains after amputation of a limb.[393] These vestigial intuitions convince some of us that inflicting the death penalty is morally necessary, even if this infliction were to do more harm to society than good. Of course there might be additional morally relevant features upon which to base the retributive intuition of death penalty as being just; the burden of proof is shifted upon the retributivist in this case to present such morally relevant features.

Turning back to the framing the innocent scenario, presumably, our intuitions of retribution regard not just punishing defectors, but rewarding rather than punishing innocent cooperators.[394] If cooperators are punished, they have diminished incentive to cooperate in the future. Research on reciprocal altruism suggests primates and human beings have developed (emotional) dispositions favorable toward cooperators and hostile toward defectors (Trivers, 1971). Evolutionary psychology explains that our moralistic

[393] A better comparison might be cross-cultural fears or repulsions, such as repulsion to spiders, as mentioned in the footnote above. Individuals in modern-day societies may still be afraid of spiders, but have little substantive reason to have such aversion: few household spiders are poisonous, and medical attention is widely available. Studies of spider phobia have suggested this fear has its origins in both evolutionary and social factors, and is significantly correlated with one's disgust response, whether genetic or learned (Seligman, 1971; Ohman 1986).

[394] Unfortunately there don't seem to be as many studies to substantiate this specific point empirically, though there are several recent studies that illustrate reciprocal altruism in monkeys as well as humans. For instance, see de Waal, F. B. M., 1997.

emotions motivate us to act favorably toward cooperators and altruists, rewarding them through social approbation, reputation, and reciprocal benefit.

Framing the innocent man scenario draws some of its argumentative force from such basic intuitions that (partially) stem from reciprocal altruism, which underlies our notion of justice. The utilitarian prescription supposed in this scenario is said to conflict with the moral prescription of these intuitions.

Generally, our intuitions regarding desert seem credible: they typically are based on morally relevant features. However, as previously illustrated in the case of retribution intuitions, we may have reason to attribute diminished credibility to such intuitions in novel contexts. The novel context, in this case, is not our current technologically-advanced environment, contrasted against our evolutionary environment, but is rather the juxtaposition of a reality-based environment versus a hypothetical, artificial environment.

The environment where the framing the innocent man scenario takes place is a novel one. And just as retributive intuitions are less credible in a novel society regulated by a sophisticated justice and penal system, in a similar way our moral judgments may be less credible in a society remarkably different from the societal context for which they were originally developed. For instance, in the present scenario, the context is one where there is epistemic certainty of consequences, and furthermore that that the certainty is that typically disutility-inducing acts will actually result in good consequences. In this way, the framing the innocent man example is asking our intuitions to extend beyond the bounds for which our intuitions are prepared.[395] Our intuitions are inculcated to deal with reality not fantasy, and hypothetical examples which stipulate epistemic certainty and fixed consequences are not realistic examples. Hare (1981) emphasizes this point:

[395]Judith Jarvis Thomson (1971) has been criticized for her reliance upon intuitions in bizarre cases which are, in turn, supposed to elucidate agential rights to abortion. In her article, "A Defense of Abortion", Thomson parallels pregnancy with several bizarre supposed analogs: for instance, having a temporarily comatose violinist attached to you, people-seeds growing in the carpet of your home, a rapidly growing child that will crush you, and so forth. The farther these thought experiments depart from reality, the less they seem credible as reference points in our moral investigations. That said, it should be noted that hypothetical examples are indeed often useful in freeing us from prejudices, or by providing a different angle of investigation. For instance, in examining whether or not its morally justified to subjugate animals for humans' carnivorous diet, we might consider whether we feel it would be justified for intellectually superior space-aliens to subject us. At least the consideration of this question might give us more clarity in our discussion, possibly freeing us from a meat-eating bias regarding our moral responsibilities to animals.

> "Perhaps the sheriff should hang the innocent man in order to prevent the riot in which there will be many deaths, if he knows that the man's innocence will never be discovered and that the bad indirect effects will not outweigh the good direct effects; but in practice he never will know this" (p. 164).

This epistemic assumption of certain knowledge of consequences seems an unrealistic assumption. Hare also mentions that it is also unrealistic to assert that the agent can be certain that the riots will be prevented by framing the innocent person, and that the riots would otherwise occur. Epistemic certainly is a subcategory of the underlying complaint that could be made by the utilitarian: that the counterexample in question is so out of the ordinary from what we could be expected to encounter, that our moral judgment of the case cannot be reasonably expected to be reliable.

It might be questioned whether this certainty that "no one will find out" is in fact unrealistic. Human agents should be assessed qua human beings, and human beings have imperfect knowledge. Yet even in a case where it's nearly certain no one will find out, the question is whether or not the agent is committing the "right action" in accordance with dispositional utilitarianism. Is a woman who is willing to frame an innocent man an agent with an optimal dispositional set? The woman, as well as the other agents involved – police, judges, lawyers, and even institutions – would seem to have to be dispositional sets that are suboptimal.

One way we can ensure the certainty stipulation is to consider an example in the past, such as looking back on "cold cases." Imagine an investigator uncovers a successfully concealed framing, where an innocent man was sacrificed for the good of others. Would the investigator of such a case, believe those actions, back then, were ethical? Our commonsense morality would condemn this uncovered conspiracy; it wouldn't change our mind that this conspiracy happened to be effective and maximized utility. Does commonsense morality in such a case, then, conflict with the prescriptions of dispositional utilitarianism? I argue that it does not. Dispositional utilitarianism would concur with commonsense morality, as effecting a conspiracy is not the "right action" for

the agents involved vis-à-vis dispositional utilitarianism.[396] The agents did not know the secret would never get out, and even if they did, a person who is ready to sacrifice justice at any point where utility would be maximized does not possess an optimal dispositional set. I will further explore the caveat of dispositional sets in relation to this example later in this section.

For the sake of argument, I will put aside this objection from boundedness, for now, and accept the epistemic certainty condition that framing the innocent man will atypically result in positive results from norm violations. From there, we might proceed by focusing on predication: namely, morally relevant features. As brought out in chapter four, intuitions prohibiting the exploitation of innocent cooperators are typically predicated upon on a preponderance of MRFs in normal contexts;[397] nevertheless, in novel contexts, there's good reason to expect that the MRFs for these intuitions may have diminished.[398] Subsequently, the credibility of our intuitions regarding this case have proportionately diminished. Our intuitions are prepared from and sensitive to reality-based cases, and having been inculcated this way, are not sensitive to the change from a reality-based context to a hypothetical one which contradicts realism. This "sleight of hand" that occurs when switching from a reality-based context to a hypothetical thought-experiment, where epistemic certainty is guaranteed and positive consequences correlate with "bad" actions, is not tracked by our intuitions and so will not change accordingly. This is a shell-game our common intuitions cannot accurately follow.

---

[396] This example seems similar to one a gambler might give. Imagine you give your brother some money to go to the store to get groceries. He passes by a casino and decides to try a spin at the roulette wheel. It turns out he gets lucky and doubles the money on a roulette spin. He comes home with the groceries and extra cash and shares his good fortune. He might justify his action of risking your shared money by saying his gambling turned out to have good consequences. However, this justification doesn't wash: even if his action just so happens to have doubled the money, it does not justify – even *post hoc* – his action. His disposition was a reckless one that endangered your shared resources; in the long run, this disposition would lead to bad consequences for you both.

[397] One morally relevant feature is the high probability exploitation of cooperators will lead to negative consequences to the integrity of the community and, thereby, the welfare its members.

[398] Consequences concerning happiness, well-being, suffering, preferences, and so forth, are morally relevant features. In addition, the *degree* of these consequences is morally relevant. In this way, if a situation changes where the negative consequences are not as severe, we can say the morally relevant features have diminished. For instance, if eating pork hardly ever results in any significant health problem for the consumer, we can say that the socially-inculcated intuition prohibiting the eating of pork was originally predicated upon morally relevant features, but in the new context these morally relevant features have diminished and, *ceteris paribus*, likewise the credibility of the intuition.

In addition to the caveat of boundedness, we can return to the first caveat concerning predication, and challenge the credibility of the moral intuitions averse to framing the innocent man. One way to do this is by diminishing the credibility of the doctrine of doing and allowing (DDA).[399] Our intuition tells us that framing one innocent person is morally worse than allowing hundreds of innocent people to die. However, if there is a sufficient nonmoral explanation for the prescriptive feeling of this intuition, this can diminish the credibility of the intuition as morally normative. In this case, a sufficient psychological explanation for DDA diminishes the credibility of the doing/allowing distinction.[400] Certainly framing the innocent man *feels* morally worse than not preventing the death of hundreds of people, but the sufficient psychological explanation regarding DDA recommends we attribute diminished credibility to this intuition.

In addition to boundedness and predication, we can invoke the second caveat concerning dispositional sets: our intuitions are likely sensitive to not just the particular instance concerning the proposed singular action in the case, but to dispositional features in general. That is, rather than predicating upon personal dispositions – such as the caring mother versus the sacrificing mother – our moral intuitions are predicating upon *institutional* dispositions: Do we endorse dispositions to act in accord with the norms of an institution of justice?[401]

An institution, such as a justice system, requires a productive dispositional set in its participants if it is to maximize utility; it has a counterproductive disposition set if it fails to do this. Presumably, as I have argued in reference to agential actions, our moral intuitions are sensitive to the dispositional set of an institution. In the framing the innocent man scenario, exploitation and massive deceit – typically undermining agents of group cohesion – atypically ensure positive results to the welfare of a community and its members. Put another way: It appears that the best way to secure the integrity of the

[399] This argument is presented in chapter 4.

[400] As explained in chapter four, it might be the case that a moral intuition, such as the doing/allowing distinction, is overdetermined; nevertheless, a sufficient causal explanation for the presence of an intuition at least diminishes the credibility of other explanations that are not in evidence. (See Joyce, 2006, chapter 6).

[401] A sacrifice of the former in the interest of the latter will indicate or lead to an erosion of the former: e.g., a mother – or institution – that can makes expedient sacrifices for the greater good, will quickly become devoid of those characteristics that make it a utility-maximizing mother or institution.

society and its members in this example is by violating the very precepts that actually ensure it under normal, real-life circumstances. Is a justice system that is readily capable of conspiring to frame an innocent man for the greater good an institution that is capable of maximizing utility in the long-run? It's doubtful. Our moral intuitions are sensitive to such dispositional sets; just as it they are sensitive to the agent in the trolley cases, or the cases where the mother need sacrifice her child. We should also note that justice systems do, while not framing the innocent, sometimes favor deterrence beyond a party's degree of guilt.[402] One example is cases of statutory rape, where say an adult male of 20 years of age is in a sexual relationship with a female of 17 years of age, who is still a minor, and cannot legally consent. The 20-year-old male might still be charged with statutory rape in order to deter cases that seem to be more blatantly egregious: after all, allowing a rule to be laxly enforced may undermine how seriously the rule is taken by those it governs.

In the framing the innocent man case, our intuitions are not limited solely to the action itself, but to the endorsement of the institution that would prioritize expediency above fairness. We might ask what kind of dispositional set the participants in an institutional system must have in order to execute such a breach of beneficent rules and processes. An institution that allowed its participants to frame an innocent person when expediency demanded it would not be a very stable or beneficial institution to its citizens. Our moral intuitions are sensitive to the MRFs regarding dispositional sets, and are not just about the act of framing an innocent citizen *ex nihilo*.

Foot (1959) makes a point relatively similar to the one we're considering: "The man who has the virtue of justice is not ready to do certain things, and if he is too easily tempted we shall say that he was ready after all" (p. 130). This sentiment also applies to institutions:[403] The institution that is beneficial to its subjects in the long-term (where stability provided by citizens is a necessary condition) would have to have participants who are not too easily disposed to perform certain actions (such as sacrificing innocent

[402] The American justice system, among others, occasionally will punish someone beyond their desert, to use them "as an example" to thereby deter violence by others.

[403] In our discussion, we're considering justice not as a virtue but, more generally, as a disposition toward beneficial treatment which favors the community, and, thereby, the individual subjects, is based on some kind of basic fairness, and respects some placeholder of "rights."

citizens for its masses), and if its participants are too easily disposed we shall say it was not a stable and beneficial institution after all.

To briefly summarize this last caveat, regarding the framing the innocent man case: The objects of our intuitions are not just the actions possible, but what institutional disposition we would be endorsing/wishing to be the case. Our intuitions lead us toward endorsing institutions that are just and uncorrupt; the reasons we endorse this kind of institution are both social and evolutionary in nature.

Socially, we may culturally endorse fairness from rational and emotional bases: we recognize it to be formally correct, as there's nothing special about us to count ourselves as an exception, and no individual member wants to be treated as expendable for the greater good.[404]

Evolutionarily, we likely have developed negative emotions toward social dynamics which punish or reward in an unjust manner. For example, in a recent study (Silk, et al., 2005), capuchin monkeys reacted negatively (and "irrationally") when they witnessed other monkeys receiving greater rewards then they did for performing the same task, even refusing subsequent food rewards in protest to the distributive inequity. If the capuchin monkey were "rational,"[405] she should take the food reward irrespective of the greater food reward given to another monkey in her group.[406] This study is similar to the "public goods game" experiments conducted in experimental economics, where human subjects contribute resources despite the fact that it is to their benefit to abstain.[407] In both studies, subjects – whether human or primate – reveal an emotional and/or moral sense of just treatment; moreover, this sense seems to be dispositional in nature, where

---

[404] This could be seen as an invocation of the golden rule, or as a possible manifestation of Kant's 1st or 2nd formulation: we could not consistently will a universal law where one treat another as a mere means to an end, as we wouldn't want ourselves to be treated in that way.

[405] The sense of "rational" here is the one used by economists, where the subject is said to be rational if she selects the option that maximizes benefit for herself.

[406] Interestingly, the capuchin receiving a less-desirable cucumber slice didn't end up taking out her anger out on her grape-receiving compatriot; instead, Brosnan reports that, based on the trajectory of the cucumber, it appeared the "blame" for the inequity was aimed at the scientists in charge of reward distribution.

[407] Either by not contributing to a common pool of resources which is subsequently doubled and then divided equally amongst the four subject, or by sacrificing some resources to punish defectors (even though they will not play with them again, or are merely observers). See Fehr and Fischbacher, 2004; Knutson, 2004; Carpenter et al., 2004.

the "rational" action is passed over in favor of the retributive action.[408] For example, in the public goods game, the human subject will spend resources to punish the human defector even if she knows she will not encounter the defector in any subsequent game. The reason for "irrational" retribution seems to be the satisfaction the subject receives from punishing the defector.[409]

In summary of the case of framing the innocent man, we can invoke the three caveats. First, one of the moral intuitions at work here regards the doing/allowing distinction: framing the innocent man is better than allowing hundreds of innocent people be killed from the violence of race riots. If the doing/allowing distinction is impugned, then the credibility of our counterintuition against the presumed utilitarian prescription is likewise diminished. Second, the boundedness caveat suggests that the hypothetical situation, as Hare emphasizes, is not the proper object of our moral intuitions when it's stipulated that utility will, with certainty, be maximized and "no one will find out." We cannot rely upon our intuitions in these cases, or at the very least their reliability is diminished. Third, the dispositional sets caveat argues that our moral intuitions are sensitive to dispositional sets of an institution, and a judicial system in which its participants feel free to frame an innocent man for the greater good is one that is counterproductive in the long-term. The defense against the framing the innocent man is twofold: (1) dispositional act utilitarianism does not endorse framing the innocent man (2) our intuitions that morally judge framing the innocent man as immoral are intuitions that suffer diminished credibility; as such, these intuitions should not be trusted. These two lines of defense cannot work in tandem, but they do both contribute to a stronger overall defense of utilitarianism against the framing the innocent man objection.[410]

---

[408] The evolutionary explanation is that the emotional and/or moral disposition is reparative to group cohesion and, thereby, evolutionarily advantageous for the individual to have . This could be consistent with a social explanation as well: for instance, how certain emotions seem reparative of a community's well-being, such as anger, guilt, shame, and so forth. This makes the action rational in a long-term way (to the individual or his genetic relatives who benefit from a stabilized group).

[409] This "satisfaction" is literal: the pleasure centers of the brain actually light up, during fMRI scans, upon punishing defectors. Human subjects seem to seek to maintain a "moral equilibrium" and are unsettled by the idea of defectors "unjustly" benefiting from transgressions and evading punishment. See Lerner, 1980; Vidmar and Miller, 1980.

[410] The point to be made here is similar to a joke told by a former professor of mine: C. Kenneth Waters. A lawyer is defending his client against charges that his dog got out of the yard and bit a passerby. The

Dissertation Conclusion

This fifth and final chapter has illustrated the upshot of my dissertation project. The method of wide reflective equilibrium, fortified with a nuanced treatment of moral judgment, provides a methodological approach that can help adjudicate moral debate. This is achieved by restructuring ethical theories, vetting moral intuitions, incorporating background theories, and adjusting all three to achieve wide coherence in reflective equilibrium.

In conclusion to this project, I offer a brief retracing of the road we've traveled. In the first chapter, I explored the current scientific understanding of moral cognition and illuminated how we arrive at our moral judgments; by elucidating how moral cognition functions, we are better situated to know how and when it might malfunction.

In the second chapter, I showed how MWRE was similar to scientific methodology, and argued that it should be acknowledged as a promising moral methodology. As part of chapter two, I responded to objections to MWRE. For instance, I responded to the objection that MWRE is circular; in its defense, I showed that the purported circularity was not vicious, and that any difficulties with the method were similar to difficulties that arise for scientific methodology, and likewise should be seen as surmountable (or at least acceptable).

In the third chapter, I articulated how moral intuitions might be determined noncredible or have diminished credibility: namely, if they are based upon morally irrelevant features, or arise under error-disposed conditions. In illustration of error-disposed conditions, I showed how empirical studies of disgust and anger, as moral emotions, can unduly influence moral judgments in certain contexts.

In chapter four, I incorporated various empirical sciences into our discussion, showing how appeal to various empirical studies could help determine, in a non-circular

---

lawyer says: "First of all, the dog couldn't have gotten out of the yard, as my client's fence is six feet high. Secondly, my client doesn't even own a dog." Though both claims cannot, together, be relevant, both claims do strengthen, together, the defense's overall case.

way, the credibility of intuitions. I focused on moral judgments regarding incest, kin preference, doing versus allowing, trust assessments, and retributive justice.

In chapter five, I put the expanded methodology of wide reflective equilibrium to work, showing how the fortified methodology could help adjudicate moral debate, with specific demonstration via utilitarian theory. In this final chapter, I illuminated three caveats of which we need be mindful in regard to our moral judgments: predication, dispositional sets, and boundedness. I showed that cognizance of these three caveats would help advance ethical debate: providing a better understanding of our moral judgments, revising our ethical theories in accommodation of empirical evidence, and understanding the limits to which we can rely upon our intuitions and principles.

The moral methodology of MWRE, as exposited in chapter two, is similar to scientific methodology. As in science, the generating conditions of data must be quality-checked to ensure data is not error-disposed. From vetted data, hypotheses and theories are tested and constructed; background theories simultaneously bear upon vetted data and hypotheses. In both scientific and moral methodologies, wide coherence determines which complex set should be accepted.

Before relying upon our moral judgments as legitimate moral data, however, we need to be mindful of intuition credibility: intuitions, for instance, must be based on morally relevant features. Also, as conveyed in this chapter, for example, moral judgments must be associated with the proper objects, such determining whether the referents of judgments are agential agents or dispositional sets. Moral intuitions must undergo a filtration process, originally proposed by Rawls, which is explored and fortified in chapter three and four. Chapter three delineates how etiologies can diminish the strength of intuitions if the etiology shows that the intuition is based, partially or fully, upon morally irrelevant features, or arose under error-disposed conditions (providing that diligent search does not locate any morally relevant features). Chapter four presents several social and evolutionary etiologies that, I argue, significantly diminish the credibility of certain categories of moral intuitions we commonly presume are reliable.

When moral theories and moral principles are put on trial, intuitions serve as members of the jury. We should be militant in quality-checking and understanding our moral judgments: what they are about, where they come from, how they can be influenced, and to what they refer. Throughout this project, I've sought to substantiate my philosophical argumentation with various empirical sciences: experimental moral psychology, evolutionary biology, cognitive science, cultural anthropology, sociology evolutionary psychology, etc. Inclusion of such empirical sciences is crucial to moral methodology, as it offers relevant background theories for greater coherence, and provides independent bases upon which to determine intuition credibility.

Moral intuitions serve as the basis of our moral navigation of the world. Empirical sciences can help investigate our moral cognition, including vetting the credibility of our moral judgments. By applying the moral methodology of wide reflective equilibrium, analogous to scientific methodology, we can make progress in moral discovery and ethical justification.

BIBLIOGRAPHY

Adams, R. M. (1976). "Motive utilitarianism." *The Journal of Philosophy*, 73, 467-481.

Alcock, J. (2003). *The Triumph of Sociobiology*. Oxford: Oxford University Press.

Alexander, R. D. (1987). *The Biology of Moral Systems*. New York: Aldine De Gruyter.

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). "The Carrot or the Stick: Rewards, Punishments, and Cooperation." *American Economic Review*, 93 (3), 893–902.

Anscombe, G. E. M. (1981). "Mr. Truman's Degree." Reprinted in *The Collected Philosophical Papers of G. E. M. Anscombe*, vol. III, Ethics, Religion and Politics, 62-71. Blackwell, Oxford.

Audi, R. (2004). *The Good in the Right*. Princeton University Press.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Bargh, J. A., Chen, M., & Burrows, L. (1996). "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action," *Journal of Personality and Social Psychology*, 71, 230-244.

Bennett, J. (1974). "The Conscience of Huckleberry Finn," *Philosophy*, 49, 123-134.

Bentham, J. (1789). *Introduction to the Principles of Morals and Legislation.* Oxford: Oxford University Press.

Blackburn, S. (1998). *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Clarendon Press.

Boyd R. (1989). "Mistakes Allow Evolutionary Stability in the Repeated Prisoner's Dilemma," *Journal of Theoretical Biology*, 136, 47-56.

Boyd, R., Bowles S., Gintis, H., & Richerson, P. (2003). "The Evolution of Altruistic Punishment," *Proceedings of the National Academy of Sciences*, 100 (6), 3531-3535.

Boyd, R., & Richerson, P. (1989). "The Evolution of Reciprocity in Sizable Groups," *Journal of Theoretical Biology*, 132, 337-356.

Boyd, R., & Richerson, P. (1992). "Punishment Allows the Evolution of Cooperation (or Anything Else) in Sizable Groups." *Ethology and Sociobiology*, 13, 171-195.

Brandt, R. (1979). *A Theory of the Good and the Right*. Oxford: Oxford University Press.

Brandt, R. (1990). "The Science of Man and Wide Reflective Equilibrium," *Ethics*, 100 (2), 259-278.
Braun C., Gruendl, M., Marberger C., & Scherber C. (2001). *Beautycheck. Causes and Consequences of Human Facial Attractiveness*. Germany: The German Students Award, Regensburg and Rostock.

Brink, D. (1987). "Rawlsian Constructivism in Moral Theory." *Canadian Journal of Philosophy*, 17 (1), 71-90.

Brody, B. (1979). "Intuitions and Objective Moral Knowledge." *Monist*, 62, 446-456.

Brosnan, S. F. (2004). "A Sense of Fairness in Monkeys." *The Encyclopedia of Animal Behavior*. M. Bekoff, ed. Greenwood Press.

Brosnan, S. F. (2006). "Nonhuman Species' Reactions to Inequity and their Implications for Fairness." *Social Justice Research*, 19 (2), 153-185.

Brosnan, S. F., Schiff, H. C.,, & de Waal, F. B. M. (2005). "Tolerance for Inequity May Increase with Social Closeness in Chimpanzees." *Proceedings of the Royal Society of London*, Series B (1560), 253-258.

Brosnan, S. F., & de Waal, F. B. M. (2002). "Variations on Tit-for-Tat: Proximate Mechanisms of Cooperation and Reciprocity." *Human Nature*, 13 (1), 129-152.

Brosnan, S. F., & de Waal, F. B. M. (2003). "Monkeys Reject Unequal Pay." *Nature*, 425, 297-299.

Brosnan, S. F., & de Waal, F. B. M. (2004a). "A Concept of Value during Experimental Exchange in Brown Capuchin Monkeys." *Folia Primatologica*, 75, 317-330.

Brosnan, S. F., & de Waal, F. B. M. (2004b). "Socially Learned Preferences for Differentially Rewarded Tokens in the Brown Capuchin Monkey." *Journal of Comparative Psychology*, 118, 133-139.

Brosnan, Sarah F., & de Waal, F. B. M. (2004c). "Reply to 'Inequity aversion in capuchins.' *Nature*, 428, 140.

Brosnan, S. F., & de Waal, F. B. M. (2005). "Responses to a Simple Barter Task in Chimpanzees, Pan Troglodytes." *Primates*, 46 (3), 173-82.

Carson, T. (1993). "Hare on Utilitarianism and Intuitive Morality." *Erkenntnis*, 39, 305-331.

Chagnon, N. A. & Bugos, P. E. (1979). “Kin Selection and Conflict: An Analysis of a Yanomamö Ax Fight.” In *Evolutionary Biology and Human Social Behavior: An Anthropological Perspective*, 86-131. N. A. Chagnon and W. Irons, eds. North Scituate, MA: Duxbury Press.

Clark, B. (2006). “Westen: Partisan Brains Can Keep Politics, Facts Separate.” *Emory Report Homepage*, 58 (23). Web.

Cosmides, L., & Tooby, J. (1992). “Cognitive Adaptations for Social Exchange.” *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, 163-221. New York: Oxford University Press.

Cummins, R. (1998). “Reflection on Reflective Equilibrium.” In *Rethinking Intuition*, 113-127. Maryland: Rowman & Littlefield Publishers, Inc.

Daly, M., Wilson, M. (1987). “Evolutionary Psychology and Family Violence.” *Sociobiology and Psychology*, in C. Crawford, M. Smith & D. Krebs, eds., Hillsdale, NJ: Erlbaum.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books.

Damasio, A. (2005). “Brain Trust,” *Nature*, 435, 571-572.

Daniels, N. (1979a). “Wide Reflective Equilibrium and Theory Acceptance in Ethics.” *The Journal of Philosophy*, 76 (5), 256-282.

Daniels, N. (1979b). “Moral Theory and the Plasticity of Persons.” *The Monist*, 62 (3), 265-287.

Daniels, N. (1980a). “On Some Methods of Ethics and Linguistics,” *Philosophical Studies*, 37, 21-36.

Daniels, N. (1980b). “Reflective Equilibrium and Archimedean Points.” *Canadian Journal of Philosophy*, 10 (1), 83-103.

Daniels, N. (1996a). “Introduction: Reflective Equilibrium in Theory and Practice.” In *Justice and Justification*, 1-20. Cambridge: Cambridge University Press.

Daniels, N. (1996b). *Justice and Justification: Reflective Equilibrium in Theory and Practice*. Cambridge: Cambridge University Press.’

Daniels, N. (1996c). “Reflective Equilibrium and Archimedean Points.” In *Justice and Justification*, 47-65. Cambridge: Cambridge University Press.

Daniels N. (1996d). “Two Approaches to Theory Acceptance in Ethics,” *Justice and Justification: Reflective Equilibrium in Theory and Practice*, 81-102. New York: Cambridge University Press.

Darley, J. M. & Latané, B. (1968). “Bystander Intervention in Emergencies: Diffusion of Responsibility.” *Journal of Personality and Social Psychology* 8, 377-383.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection.* London: Murray.

Darwin, C. (1874). *The Descent of Man, and Selection in Relation to Sex,* 2nd ed. London: Murray.

Dawkins, R. (1976). *The Selfish Gene.* Oxford: Oxford University Press.

De Waal, F. B. M. (1996). *Good Natured: The Origin of Right and Wrong in Humans and Other*
Animals. Cambridge, MA: Harvard University Press.

De Waal, F. B. M. (1997). “Food Transfers through Mesh in Brown Capuchins.” *Journal of Comparative Psychology*, 111, 370-378.

Dennett, D. (1996). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon & Schuster.

DePaul, M. (1986). “Reflective Equilibrium and Foundationalism,” *American Philosophical Quarterly*, 23 (1), 59-69.

DePaul, M. (1998). “Why Bother with Reflective Equilibrium?” In *Rethinking Intuition*, 293-309. Maryland: Rowman & Littlefield Publishers, Inc.

DePaul, M. R., & Ramsey, W., eds. (1998). *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*. Lanham: Rowman & Littlefield Publishers.

Descartes, R. (1641). *Meditations on First Philosophy.* Cottingham, J., trans. (1996). Cambridge University Press. Latin original.

Devine, P. G. (1989). “Stereotypes and Prejudice: Their Automatic and Controlled Components.” *Journal of Personality and Social Psychology*, 56, 5-18.

Dion, K., Berscheid, E., & Hatfield, E. (1972). “What is Beautiful is Good.” *Journal of Personality and Social Psychology*, 24, 285-290.

Ditto, P. H., & Lopez, D. F. (1992). “Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Non-Preferred Conclusions.” *Journal of Personality and Social Psychology*, 63, 568-584.

Duhem, P. (1954). *The Aim and Structure of Physical Theory*, Princeton University Press. Translated from the French by Philip P. Wiener.

Edward, S. (1985). “Wide Reflective Equilibrium and Science.” *Southwest Philosophy Review*, 2, 105-115.

Epley, N., & Dunning, D. (2000). “Feeling ‘holier than thou’”. *Journal of Personality and Social Psychology*, 79, 861-875.

Fehr E., & Gachter, S. (2002). “Altruistic Punishment in Humans.” *Nature*, 415: 137-140.

Flack, J. Girvan, M., de Waal, F., & Krakauer, D. (2006). “Policing Stabilizes Construction of Social Niches in Primates.” *Nature*, 439: 426-429.

Flinn, M. V., Leone, D. V., and Quinlan, R. J. (1999). “Growth and Fluctuating Asymmetry of Stepchildren.” *Evolution and Human Behavior*, 20, 465-479.

Foot, P. (1958). “Moral Arguments.” *Mind*, 67 (268), 502-513.

Foot, P. (1959). “Moral Beliefs.” *Proceedings of the Aristotelian Society (1958-59)*, 59, 83-104.

Foot, P. (1967). “The Problem of Abortion and the Doctrine of Double Effect,” *Oxford Review*, 5, 5-15.

Foot, P. (1977). *Virtues and Vices.* Oxford: Blackwell.

Foot, P. (1984). “Killing and Letting Die,” *Abortion: Moral and Legal Perspectives,* 177-185. Garfield, J., ed. Amherst: University of Massachusetts Press.

Gibbard, A. (1982). “Human Evolution and the Sense of Justice.” In French, *Social and Political Philosophy,* 31-46.

Gibbard, A. (1990). *Wise Choices, Apt Feelings*. Cambridge, Harvard University Press.

Giere, R. (1999). “Using Models to Represent Reality.” *Model-Based Reasoning in Scientific Discovery*, 41-57. L. Magnani, N. J. Nersessian, and P. Thagard, eds. New York: Kluwer/Plenum.

Gilligan, C. (1977). “In a Different Voice: Women’s Conception of Self and Morality.” *Harvard Educational Review*, 47, 481-517.

Gladwell, M. (2005). *Blink: The Power of Thinking without Thinking*. New York: Little Brown & Co.

Glover, J. (1977). *Causing Death and Saving Lives*. Harmondsworth, Penguin Books.

Greene, J. D. (2002). *The Terrible, Horrible, No Good, Very Bad Truth About Morality, and What to Do About It.* (Ph.D. dissertation, Department of Philosophy, Princeton University), Chapter 3. (advised by Lewis, D., and Harman, G.)

Greene, J.D. (2003). “From Neural ‘is’ to Moral ‘ought’: What are the Moral Implications of Neuroscientific Moral Psychology?” *Nature Reviews Neuroscience*, 4, 847-850.

Greene, J. D. (2007). “The Secret Joke of Kant's Soul,” *Moral Psychology* 3*: The Neuroscience of Morality*: *Emotion, Disease, and Development*, 35-80. W. Sinnott-Armstrong, ed. MIT Press: Cambridge, MA.

Greene, J.D., Baron, J. (2001). “Intuitions about Declining Marginal Utility.” *Journal of Behavioral Decision Making*, 14, 243-255.

Greene, J. D., & Haidt, J. (2002). “How (and Where) Does Moral Judgment Work?” *Trends in Cognitive Sciences*, 6, 517-523.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). “An fMRI Investigation of Emotional Engagement in Moral Judgment.” *Science*, 293, 2105-2108.

Hacking, I. (1983), *Representing and Intervening.* New York: Cambridge University Press.

Haidt, J. (2000). “The Moral Emotions,” *Handbook of Affective Sciences*, 852-870. R. J. Davidson, K. Scherer and H. H. Goldsmith, eds. New York: Oxford University Press.

Haidt, J. (2001). “The Emotional Dog and Its Rational Tail.” *Psychological Review*, 198 (4), 814-834.

Haidt, J. (2003). “The Emotional Dog Learns New Tricks.” *Psychological Review*, 110 (1), 197–198.

Haidt, J., Bjorklund, F., & Murphy, S. (2004). *Moral Dumbfounding: When Intuition Finds No Reason*. (Unpublished manuscript, University of Virginia).

Haidt, J., Koller, S. H., & Dias, M. G. (1993). "Affect, Culture, and Morality, or is it Wrong to Eat your Dog?" *Journal of Personality and Social Psychology*, 65 (4), 613-628.

Hanson, R. (2002). "Why Health is Not Special: Errors in Evolved Bioethics Intuitions." *Social Philosophy & Policy*, 19 (2), 153-179.

Hare, R. M. (1981). *Moral Thinking: Its Level, Method, and Point*. Oxford: Clarendon.

Hare, R. M. (1989). *Essays in Ethical Theory*. Oxford: Clarendon.

Harman, G. (1975). "Moral Relativism Defended." *The Philosophical Review*, 84 (1), 3-22.

Harman, G. (1977). *The Nature of Morality*. New York: Oxford.

Harman, G. (1986). "Moral Explanations of Natural Facts – Can Moral Claims Be Tested Against Reality?" *The Southern Journal of Philosophy*, 24 (supplement), 69-78.

Harman, G. (2003). "Three Trends in Moral and Political Philosophy." *The Journal of Value Inquiry*, 37 (3), 415-425.
Harms, W. (2000). "Adaptation and Moral Realism." *Biology and Philosophy*, 15, 713-732.

Hauser, M. D. (2005). "Moral Ingredients: How We Evolved the Capacity to Do the Right Thing," *Evolution and Culture: A Fyssen Foundation Symposium*. Stephen C. Levinson and Pierre Jaisson, eds. MIT Press, 219-246.

Henrich J, & Boyd R (2001). "Why People Punish Defectors." *Journal of Theoretical Biology*, 208, 79–89.

Herrnstein, R. & Murray, C. (1994). *The Bell Curve.* New York: The Free Press.

Holmgren, M. (1987). "Wide Reflective Equilibrium and Objective Moral Truth." *Metaphilosophy*, 18, 108-125.

Holmgren, M. (1989). "The Wide and Narrow of Reflective Equilibrium," *Canadian Journal of Philosophy*, 19 (1), 43-60.

Horowitz, T. (1998). "Philosophical Intuitions and Psychological Theory," *Rethinking Intuition,* 143-159. Maryland: Rowman & Littlefield Publishers, Inc.

Hughes, W. (1986). "Richard's Defense of Evolutionary Ethics." *Biology and Philosophy*, 1, 306-315.

Hume, D. (1742). *Essays, Moral and Political*. Edinburgh: A. Kincaid.

Hume, D. (1751). *An Enquiry Concerning the Principles of Morals.* London: A. Millar.

Hume, D. (1964). *A Treatise of Human Nature.* L. A. Selby-Bigge, ed. Oxford: Clarendon.

Hume, D. (1983). *An Enquiry Concerning the Principles of Morals*. Indianapolis: Hackett.

Hume, D. (1987). *Essays, Moral, Political, and Literary* vol. 1, 91-266. Indianapolis: Liberty Fund, Inc.

Hursthouse, R. (1999). *Virtue Ethics*. New York: Oxford University Press.

Isaac, R., Walker, J., and Williams, A. (1994). "Group Size and the Voluntary Provision of Public Goods: Experimental Evidence Utilizing Large Groups." *Journal of Public Economics*, 54 (1), 1-36.

Janssen, M. and Ahn, T. (2003). "Adaptation vs. Anticipation in Public-Good Games." *American Political Science Association meetings*, Philadelphia, PA., August 27.

Johnson, O. (1957): "Ethical Intuitionism – A Restatement." *The Philosophical Quarterly*, 7 (28), 193-203.

Johnson, R. (2003). "Virtue and Right." *Ethics*, 113, 810-834.

Joyce, R. (2000). "Darwinian Ethics and Error." *Biology and Philosophy*, 15 (5), 713-732.

Joyce, R. (2006). *The Evolution of Morality.* London: MIT Press.

Kamm, F. M. (1998). "Moral Intuitions, Cognitive Psychology and the Harming-versus-Not-Aiding Distinction." *Ethics*, 108 (3), 463-488.

Kant, I. (1889). *Kant's Critique of Practical Reason and Other Works on the Theory of Ethics.* T.Abbott, trans., 4th revised ed. London: Kongmans, Green and Co.

Kant, I. (1981). "On the Supposed Right to Lie because of Philanthropic Concerns." *Grounding for the Metaphysics of Morals*, 3rd ed. 63-67.

Kant, I. (1998). *Critique of Pure Reason*. P. Gruyer and A. W. Wood, eds. and trans. Cambridge: Cambridge University Press.

Kant, I. (2000). "Lectures on Ethics." *The Cambridge Edition of the Works of Immanuel Kant*. Peter Heath and J. B., eds. Schneewind.. New York: Cambridge University Press.

Kant, I. (2002). *Groundwork for the Metaphysics of Morals.* A. W. Wood, trans. New Haven: Yale University Press.

Kitcher, P. (1984). *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. Cambridge, Massachusetts: Bradford Books, MIT Press.

Kitcher, P. (1986). "The Transformation of Human Sociobiology." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 2 (Symposia and Invited Papers), 63-74. Chicago: University of Chicago Press.

Kitcher, P. (1993). "The Evolution of Human Altruism." *The Journal of Philosophy*, 90, 497-516.

Kitcher, P. (1997). "Four Ways of 'Biologicizing Ethics.'" Eliot Sober, *Conceptual Issues in Evolutionary Biology*, 439-450. E. Sober, ed.. Cambridge: MIT Press.

Kitcher, P. (1998a). "Believing Where We Cannot Prove" Kitcher, *Abusing Science: The Case Against Creationism*. Cambridge: The MIT Press, 1982, pp. 30-54.

Kitcher, P. (1998b). "Psychological Altruism, Evolutionary Origins, and Moral Rules." *Philosophical Studies*, 89, 283-316.

Knobe, J. (2005). "Ordinary Ethical Reasoning and the Ideal of 'Being Yourself'." *Philosophical Psychology*, 18 (3), 327-340.

Kohlberg, L., & Turiel, E. (1971). "Moral Development and Moral Education." G. Lesser, ed. *Psychology and Educational Practice.* Glenview, Ill.: Scott Foresman.

Kohlberg, L., Levine, C., & Hewer, A. (1983). *Moral Stages: A Current Formulation and a Response to Critics*. New York: Karger.

Kosfeld, M., Heinrichs, M. Zak, P. J., Fischbacher, U., & Fehr, E. (2005). "Oxytocin Increases Trust in Humans." *Nature* 435, 673-676.

Korsgaard, C. (1996a). *Creating the Kingdom of Ends*. Cambridge University Press.

Korsgaard, C. (1996b). "The Right to Lie: Kant on Dealing with Evil" in *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press, 133-158.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lakatos, I. (1970). *Criticism and the Growth of Knowledge*. New York: Cambridge University Press.

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes: Philosophical Papers Volume 1*. Cambridge: Cambridge University Press

Lerner, M.J. (1980). *The Belief in a Just World: A Fundamental Delusion*. New York: Plenum Press.

Lewontin, R. (1972) "The Apportionment of Human Diversity." *Evolutionary Biology*, 6, 391-398.

Little, D. (1984). "Reflective Equilibrium and Justification." *Southern Journal of Philosophy*, 22, 373-388.

Littlefield, C., & Rushton, J. (1986). "When a Child Dies: The Sociobiology of Bereavement." *Journal of Personality and Social Psychology*, 51: 797-802.

Marlowe, F. (1999). "Male Care and Mating Effort among Hadza Foragers." *Behavioral Ecology and Sociobiology*, 46, 57-64.

Mason, M. (2003). "Contempt as a Moral Attitude." *Ethics*, 113, 234-272.

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

McCloskey, H. J. (1965). "A Non-Utilitarian Approach to Punishment." *Inquiry*, 8, 239-255.

Milgram, S. (1963). "Behavioral Study of Obedience". *Journal of Abnormal and Social Psychology*, 67, 371–378.

Mill, J. S. (1998). *Utilitarianism*. Roger Crisp, ed. Oxford: Oxford University Press

Miller, D. T, & Ratner, R. K. (1998). "The Disparity between the Actual and Assumed Power of Self-Interest." *Journal of Personality and Social Psychology*, 74, 53-62

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

Nagel, T. (1974). "What is it Like to Be a Bat?" *Philosophical Review*, 435-450.

Nagel, T. (1979). "Moral Luck." *Mortal Questions*. Cambridge: Cambridge University Press, 24-38.

Nielsen, K. (1977). "Our Considered Judgments," *Ratio*, 19 (1), 39-46.

Nichols, S. (2002). "Norms with Feeling: Towards a Psychological Account of Moral Judgment." *Cognition*, 84, 221-36.

Nichols, S. (2004). *Sentmental Rules: On the Natural Foundations of Moral Judgment*. Oxford University Press.

Nichols, S. (2005). "Innateness and Moral Psychology," *The Innate Mind: Structure and Contents*, 353-430. Peter Carruthers, Stephen Laurence, and Stephen Stich, eds. Oxford University Press.

Panchanathan, K. (2004). "Human Cooperation: Second-Order Free-Riding Problem Solved?" *Nature*, 437 (7058).

Parfit, D. (1973). "Later Selves and Moral Principles," *Philosophy and Personal Relations*, 137-169. A. Montefiore, ed. Routledge and Kegan Paul: London.

Parfit, D. (1987). "Divided Minds and the Nature of Persons," *Mindwaves*, 19-26. Blakemore & Greenfield eds. Oxford: Basil Blackwell.

Perlmutter, M. (1999). "Desert and Capital Punishment." In *Morality and Moral Controversies.* New Jersey: Prentice-Hall Press.

Pizarro, D. A., & Paul Bloom. (2003). "The Intelligence of the Moral Intuitions: A Reply to Haidt. *Psychological Review*, 110, 193–196.

Prichard, H.A. (1912). "Does Moral Philosophy Rest on a Mistake?" *Mind*, 21, 21-37.

Prinz, J. (2006). *The Emotional Construction of Morals*. Oxford University Press.

Putnam, H. (1978). *Meaning and the Moral Sciences*. London: Routledge & Kegan Paul.

Quinn, W. (1989). "Actions, Intentions, Consequences: The Doctrine of Doing and Allowing." *Philosophical Review*, 98, 287-312.

Quinn, W. (1993). *Morality and Action*. New York: Cambridge University Press.

Rachels, J. (1986). *The Elements of Moral Philosophy.* New York: Random House.

Railton, P. (1984). "Alienation, Consequentialism, and the Demands of Morality." *Philosophy and Public Affairs*, 13, 134-71.

Railton, P. (1986). "Moral Realism." *The Philosophical Review*, 95 (2), 163-207.

Rawls, J. (1951). "Outline of a Decision Procedure in Ethics." *The Philosophical Review*, 60 (2), 177-197.

Rawls, J. (1971). *A Theory of Justice.* Cambridge, Massachusetts: Harvard University Press.

Rawls, J. (1974-1975). “The Independence of Moral Theory,” *Proceedings and Addresses of the American Philosophical Association*, 47, 5-22.

Regan, T. (1983). *The Case for Animal Rights*. Berkeley: University of California Press.

Richards, R. (1986). “A Defense of Evolutionary Ethics.” *Biology and Philosophy*, 1, 265-293.

Richards, R. (1989). *Darwin and Evolutionary Theories of Mind and Behavior*. University of Chicago Press.

Rosenberg, A. (1991). “The Biological Justification of Ethics: A Best Case Scenario.” *Social Policy and Philosophy*, 8, 86-101.

Roojen, M. (1999). “Reflective Moral Equilibrium and Psychological Theory.” *Ethics*, 109 (4), 846-857.

Ross, W. D. (1930). *The Right and the Good*. Oxford: Clarendon Press.

Rozin, P., Haidt, J., & McCauley, C. R. (2000). “Disgust.” In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions*, 2nd edition, 637-653. New York: Guilford Press.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). “The Moral-Emotion Triad Hypothesis.” *Journal of Personality and Social Psychology*, 76, 574-586.

Ruse, M., & Wilson, E. O. (1985). “The Evolution of Ethics.” *New Scientist*, 17, 50-52.

Ruse, M., & Wilson, E. O. (1986). “Moral Philosophy as Applied Science: A Darwinian Approach to the Foundations of Ethics.” *Philosophy*, 61, 173-92.

Sahlin, M. (1965). “On the Sociology of Primitive Exchange.” *The Relevance of Models for Social Anthropology*, M. Banton, ed. London: Monographs of the A.S.A. (1), 139-236.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Scoccia, D. (1990). “Utilitarianism, Sociobiology, and the Limits of Benevolence.” *The Journal of Philosophy*, 87 (7), 329-345.

Segal, N, & Hershberger, S. (1999). “Cooperation and Competition Between Twins: Findings from a Prisoner's Dilemma Game.” *Evolution and Human Behavior*, 20, 29-51.

Sencerz, S. (1986). “Moral Intuitions and Justification in Ethics,” *Philosophical Studies*, 50 (1), 77-95.

Shaw, W. H. (1980). “Intuition and Moral Philosophy.” *American Philosophical Quarterly*, 17 (2), 127-134.

Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., Lambeth, S. P., Mascaro, J., & Schapiro, S. J. (2005). “Chimpanzees are Indifferent to the Welfare of Unrelated Group Members.” *Nature*, 437 (7063), 1357-1359.

Sidgwick, H. (1907). *Methods of Ethics*, Macmillan and Company, Ltd.

Sigmund, K., Fehr, E., & Nowak, A. (2002). “The Economics of Fair Play.” *Scientific American*, 286 (1), 82-87.

Singer, P. (1972). “Famine, Affluence, and Morality.” *Philosophy and Public Affairs*, 229-243.

Singer, P. (1974). “Sidgwick and Reflective Equilibrium.” *The Monist*, 58, 490-517.

Singer, P. (1975). *Animal Liberation*. New York: Random House.

Singer, P. (1981). *The Expanding Circle: Ethics and Sociobiology*. Oxford: Clarendon.

Singer. P. (1982). “Ethics and Sociobiology.” *Philosophy and Public Affairs*, 11 (1), 40-64.

Singer, P. (1984). “Ethics and Sociobiology.” *Zygon*, 19, 141-158.

Singer, P. (1993a). *Practical Ethics*, 2nd ed. Cambridge: Cambridge University Press.

Singer, P. (1993b). “Rich and Poor.” *Practical Ethics*, 2nd ed. Cambridge: Cambridge University Press, 218-246.

Singer, P. (2005). “Ethics and Intuitions.” *Journal of Ethics* 9, 331-352.

Singleton, J. (1981). “Moral Theories and Tests of Adequacy.” *The Philosophical Quarterly*, 31, 31-46.

Sinnott-Armstrong, W. (2008). “Framing Moral Intuitions” in *Moral Psychology* 2. Cambridge: MIT Press, 47-76.

Smart, J. J. C., & Williams, B. (1973). *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Smith, M. B. E. (1979). “Ethical Intuitionism and Naturalism: A Reconciliation,” *Canadian Journal of Philosophy*, 9, 609-629.

Sober, E., & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, Massachusetts: Harvard University Press.

Sommers, T. (August, 2005). “Interview with Jonathan Haidt.” *The Believer*.

Spielberg, S., Zaillian, S., Neeson, L., Kingsley, B., Fiennes, R., & Keneally, T. (1994). *Schindler's list*. Universal City, Calif: MCA Universal Home Video.

Stent, G. (1978). *Morality as a Biological Phenomenon*. Berkeley and Los Angeles: University of California Press.

Sturgeon, N. (1982). “Brandt’s Moral Empiricism,” *The Philosophical Review*, 91, 389-422.

Sturgeon, N. (1986). “Harman on Moral Explanations of Natural Facts,” *The Southern Journal of Philosophy*, 24 (supplement), 69-78.

Sturgeon, N. (1992). “Nonmoral Explanations,” *Philosophical Perspectives*, 6: Ethics. James Tomberlin, ed., 97-117. Atascadero: Ridgeview.

Thomson, J. (1971). “A Defense of Abortion.” *Philosophy and Public Affairs* 1 (1), 47-66.

Trivers, R. (1971). “The Evolution of Reciprocal Altruism.” *The Quarterly Review of Biology*, 46, 35-57.

Turiel, E., Hildebrandt, C., & Wainryb, C. (1991). *Judging Social Issues: Difficulties, Inconsistencies, and Consistencies*. Chicago: University of Chicago Press.

Tversky, A., & Kahneman, D. (1981). “The Framing of Decisions and the Psychology of Choice.” *Science*, 211, 453-458.

Twain, M. (1885). *The Adventures of Huckleberry Finn*. New York: Charles Webster.

Unger, P. (1996). *Living High and Letting Die*. New York: Oxford University Press.

Vidmar, N., & Miller, D. (1980). “Sociopsychological Processes Underlying Attitudes Toward Legal Punishment.” *Law & Society Review*, 14, 565-602.

Walker, M. U. (1998). *Moral Understandings*. New York: Routledge.

Westen, D., Kilts, C., Blagov, P., Harenski, K., & Hamann, S. (2006). "The Neural Basis of Motivated Reasoning: An fMRI Study of Emotional Constraints on Political Judgment during the U.S. Presidential Election of 2004." *Journal of Cognitive Neuroscience*, 18, 1947-1958.

Wheatley, T., & Haidt, J. (2005). "Hypnotically Induced Disgust Makes Moral Judgments More Severe." *Psychological Science*, 16, 780-784.

Williams, B. (1973). "A Critique of Utilitarianism," in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against.* Cambridge: Cambridge University Press.

Williams, B. (1981). "Utilitarianism and Moral Self-Indulgence." *Moral Luck: Philosophical Papers 1973-1980.* Cambridge: Cambridge University Press, 40-53.

Wilson, E. O. (1975). *Sociobiology: The New Synthesis*. Cambridge: Harvard University Press.

Wilson, E. O. (1998). *Consilience: The Unity of Knowledge.* New York: Alfred A. Knopf, Inc.

Wolf, S. (2004). "The Moral of Moral Luck." *Philosophic Exchange*; reprinted in Cheshire Calhoun, ed., *Setting One's Moral Compass: Essays by Women Philosophers*. New York: Oxford, 113-127.

Wright, L. (1973). "Functions." *Philosophical Review*, 82 (2), 139-168.

Wright, R. (1994). *The Moral Animal: The New Science of Evolutionary Psychology*. New York: Pantheon Books.