UNIVERSITY OF CALIFORNIA,

IRVINE

Forming our Moral Selves:

How Science Can Help Us Live Up to Our Moral Values

DISSERTATION

submitted in partial satisfaction of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Jessica Marie Gonzalez

Dissertation Committee:
Professor P. Kyle Stanford, Chair
Professor Cailin O'Connor
Associate Professor Lauren Ross
Professor Kristen R. Monroe
Assistant Professor Nadia Chernyak

2023

DEDICATION

This dissertation is dedicated to my children, my husband, and my parents.

Their support and encouragement made it possible to realize my dream,

and for that I am eternally grateful.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

ACKNOWLEDGEMENTS

I am fortunate that my graduate experience has included an extraordinary list of mentors. This endeavor would not have been possible without my committee chair, Professor P. Kyle Stanford. His support, enthusiasm, and guidance led me from having a scattered but intense interest in moral cognition to discover the exact philosophical question that drives me. I am forever grateful for his patience and foresight.

I am also incredibly thankful for my committee members. I thank Professor Cailin O'Connor, whose work has inspired key aspects of this dissertation and who has been instrumental in advising me through pivotal times in my graduate experience. I thank Associate Professor Lauren Ross, who taught me to think carefully and intentionally about matters of causation and biology, and who guided my growth as an academic. I thank Assistant Professor Nadia Chernyak for her exceptional hospitality in welcoming me into her developmental psychology lab, which allowed me see the field from the inside; this was crucial to the authenticity of what I wanted my project to be.

I would also like to thank Professor Barbara Sarnecka for her mentorship and support; without her writing workshops, this dissertation may very well never have come to fruition. Because of her, I am a better writer and a better mentor to others.

Lastly, and with deep appreciation, I thank Professor Kristen Monroe and the UCI Ethics Center. Nothing is more central to my project than putting it into action, and Professor Monroe created countless opportunities for me to do just this. She has been a source of continuous intellectual and moral support.

VITA

Jessica Marie Gonzalez

2005	B.A. in Political Science and Philosophy, University of California, Los Angeles
2009	M.A. in Philosophy, California State University, Los Angeles
2013-17	Instructor of Philosophy, Tenure-Track, Hawai‘i Community College
2021	M.A. in Philosophy, University of California, Irvine
2023	Instructor of Philosophy, Tenure-Track, Orange Coast College
2023	Ph.D. in Philosophy, University of California, Irvine

FIELD OF STUDY

Logic and Philosophy of Science, with an emphasis in Biological and Behavioral Sciences

ABSTRACT OF THE DISSERTATION

Forming Our Moral Selves:

How Science Can Help Us Live Up to Our Moral Values

by

Jessica Marie Gonzalez

Doctor of Philosophy in Philosophy

University of California, Irvine, 2023

Professor P. Kyle Stanford, Chair

What is the role of science in morality? Some have proposed that science can tell us what we should value. I reject this notion, showing that it leads to the naturalistic fallacy. Instead of normative value, I argue, science has practical value for our moral deliberations. We should use science to understand the cognitive processes behind our moral behaviors. This *informed introspection* will allow us to see our own moral activities descriptively, exposing the ways in which they veer away from how we prescriptively conceptualize our own moral codes. Our informed introspection will illuminate our moral beliefs, intuitions, and judgements, pointing out incoherence between our moral behaviors and our moral principles. That is, at times we may unintentionally act in ways that are contrary to our deeply held moral principles. By incorporating scientific information about our moral processes into our moral deliberations, we will gain better agency over our moral behaviors. We will know when our moral processes are likely to pull us away from our moral principles, and we will have practical ways to correct for it. To illustrate how we can use science to make concrete changes in our moral deliberations, I

provide three examples. First, an informed introspection reframes the way we think about moral regret. Second, it calls us to be skeptical of how we use factual beliefs in our moral deliberations. Third, it also calls for skepticism about our moral intuitions. Adopting these changes will help us build coherence between our moral behaviors and our moral principles, giving us more autonomy as moral agents and helping us make decisions we will be more satisfied with upon reflection.

CHAPTER 1: Finding Our Place Between Science and Morality

1.0 Introduction

> Many of our most serious conflicts are conflicts within ourselves. Those who suppose their judgments are always consistent are unreflective or dogmatic; not uncommonly they are ideologues and zealots.
>
> -John Rawls
> *Justice as Fairness: A Restatement* 2001

We are not morally consistent creatures. Sometimes we act in ways that defy our deepest held moral commitments. John Rawls hints at this when he writes that not only do judgments differ from one person to another, but we are often of a divided mind regarding our own judgments (Rawls, 2001, p. 30). If we are to build coherence between our own vying judgments, we must reflect on our moral principles, revising our judgments as needed. Philosophers often point to introspection as a way to serve this end. What I present in this dissertation illuminates a new introspection – an *informed* introspection built on a scientific understanding of moral cognition.[1] Rather than balancing only our thoughts when deliberating about a moral problem, our informed introspection incorporates what science tells us about our moral cognitive processes. By taking a fuller account of our moral intuitions, *viz*., examining their origins, we can make judgments that are more consistent with what we deeply value.

Scientists have since bequeathed us with a wealth of knowledge about our moral cognition – the ways in which our moral behavior is influenced by our cognitive processes. What these studies tell us time and time again, is that our moral thinking and our moral actions are not

[1] This idea of an *informed introspection* is inspired by Emily Pronin's (2009) concept of *introspection illusion*. I briefly expand on Pronin's concept in Chapter 3.

always coherent. Our moral actions are often influenced by factors ranging from our social environment, e.g., how many people are in the room with us (Latané & Rodin, 1969), to our neural activity, e.g., the size of our amygdala (Marsh et al., 2014). Our moral commitments, on the other hand, can be the product of a lifetime of cultural input and introspection. The incoherence between these two aspects of our moral selves is not just surprising; it can also be disconcerting.

In this introductory chapter, I provide a foundation for my claim that scientific knowledge about how our moral cognition works is important to our moral deliberations. The insight provided by science can help us build coherence between our moral thinking and our moral actions, making us more satisfied moral agents upon reflection. In other words, if we let science teach us how we might be likely to act in certain cases, we can use that information as we reflect on the moral situations we encounter. By keeping our eyes open to our moral tendencies and intuitions, we can work to understand and align our actions with the moral commitments we hold.

The approach I take here differs from other views about the role science should play in our moral thinking. Here in Chapter 1, I examine Kitcher's *Four Ways of 'Biologicizing' Ethics* in which he creates a map for would-be sociobiological ethicists so that they may find their place in the conversation and identify questions they may answer. I find that Kitcher's map is outdated, and that the advancements of moral cognition in recent decades should cause us to expand our thinking about the relationship between science and morality. As an alternative, I begin to lay the foundation for my view that science should play a practical role in our moral thinking, guiding our moral deliberations. Next, I preview the types of concrete changes I believe science can help us make. Then I connect my argument to Christine Korsgaard's concept of normative self-

government, proposing that using science in this way will make us better moral agents. By building into our awareness of our cognitive processes into our moral considerations, we will be better able to act according to our moral principles, giving us more autonomy over our moral actions.

In Chapter 2, I consider a popular approach to the question of how science should be used in ethics: the idea that science can tell us what we should value. Kitcher warns us about this project in his *Four Ways* (2006) chapter. However, in subsequent work (Kitcher, 2011), he does this very thing, arguing that scientific facts can give us normative consequences. Here, I also point to the work of Joshua Greene (2014), showing that he follows suit, claiming normative consequences from a descriptive theory. Ultimately, although there is much to learn from their empirical accounts, they each commit the naturalistic fallacy. I use Kitcher (2011) and Greene's work as a cautionary tale, showing the benefit of striving for practical over normative consequences from our scientific knowledge.

I begin Chapter 3 by providing the empirical context supporting my introspective proposal. There is a lingering question that must be addressed before we can take my proposal seriously: will it work? Although this empirical question is best answered after implementing the changes I recommend, there is research from psychology that speaks more generally about the power of introspection. I address this research, finding that it gives us reason for optimism that the introspective process I recommend will be effective.

The rest of Chapter 3 and Chapter 4 are dedicated to demonstrating the practical potential of an informed introspection. I offer three cases as a proof of concept, showing how science can make concrete changes to our moral deliberations. I dive into studies from neuroscience, cognitive science, and developmental psychology, finding that they illuminate our moral

processes by telling us surprising information about the cognitive processes underneath our moral activities. In Chapter 3, I propose that we can use this information to provide relief from moral regret in some cases. I further propose that learning about the relationship between factual and moral beliefs can help us achieve more honest moral deliberations. In Chapter 4, I argue that by learning about how we form our moral intuitions, we can work to increase harmony between our moral actions and our moral principles.

Finally, in Chapter 5, I consider the moral theories of two more philosophers, John Dewey and John Rawls. I view my account as being in accordance with that of Dewey, who sees much value in the process of moral reflection and also finds a scientific approach to morality valuable for its practical consequences. In considering Rawls's account, I look to his notion of reflective equilibrium, proposing that it would be improved by adding a scientific understanding of our moral processes, i.e., my view here. To conclude this final chapter, I return to the idea of moral agency, finding that increasing our autonomy over our moral activities will make us more satisfied agents upon reflection.

1.1 'Biologicizing' Ethics

In the previous section, I mentioned two naturalistic accounts that claim we can derive normative consequences from a scientific understanding of morality, Kitcher (2011) and Greene (2014). An arguably stronger claim was made by E.O. Wilson (1975) when he wrote that we should consider "the possibility that the time has come for ethics to be removed temporarily from

the hands of the philosophers and biologicized" (Wilson, 1975, p. 562; from Kitcher 2006).[2] Wilson's claim reflects a desire to pause nonscientific approaches to morality in order to develop scientific approaches that can guide our investigations into morality. Kitcher, in a 2006 chapter titled *Four Ways of 'Biologicizing' Ethics*, responds to Wilson's claim by teasing out the different ways in which scientific fields can contribute to our understanding of ethics.

I begin this section by presenting a summary of Kitcher's argument, wherein I review the four ways he proposes we might "biologicize" ethics. Here, I argue that Kitcher's picture is incomplete – there is another meaningful way in which scientific fields may contribute to ethics. I introduce my novel argument that scientific information matters for ethics because it can help us make concrete changes to our moral deliberations that will benefit us in the long term. In Section 2, I preview three examples of concrete changes, each of which will be more fully explored in this dissertation. Section 3 connects my view to another, established ethical view that is in harmony with my picture of the role science should play in our understanding of ethics.

In writing about the ways to biologicize ethics, Kitcher has two goals. First, he aims to organize projects that surface in work by Wilson and his co-authors, mathematical physicist Charles Lumsden and philosopher Michael Ruse. As he presents these projects, Kitcher offers an analysis of their strengths and weaknesses. This leads into his second goal, which is to provide a map for would-be sociobiological ethicists so that they may consider where their interests lie and what questions they should consider in their work (Kitcher, 2006, p. 575).

[2] In this section, I use the terms 'biologicize' (in its various forms) and 'sociobiology' to be consistent with what I refer to as 'scientific' in the rest of this dissertation. My argument applies broadly to any scientific field that investigates ethical topics. The arguments referenced in this section by Kitcher, Wilson and others, use these more specific terms, and so I use those terms here as well, but it should be noted that their use is consistent with the general terms I use in my argument. In other words, anything I say about 'sociobiology' should be read as also applying to any scientific field that investigates ethical topics.

PROJECT 1: *Sociobiology has the task of explaining how people have come to acquire ethical concepts, to make ethical judgments about themselves and others, and to formulate systems of ethical principles* (Kitcher, 2006, p. 576).

In this first project, science investigates the origins of ethical thinking in a purely descriptive manner. This project is concerned with tracing our human history far back so that we may learn more about the coevolution of our genes and culture. A possible benefit of this project, Kitcher suggests, is that it may lead us to discover that our ethical rules are not purely a product of natural selection, as neo-Darwinists hold. Instead, we may find that what natural selection has endowed us with is very general capacities for learning, and that our formulation of ethical rules is more heavily dependent on cultural selection. By exploring the coevolution of genes and culture, this project can provide a broad descriptive picture of the origin of our ethical concepts.

Kitcher is largely in favor of this project but does offer a warning: the findings here should not be overinterpreted – and they are overinterpreted, Kitcher claims, by Wilson. Wilson and Ruse combine the evolutionary history of our ability to make moral judgments with the claim that everything about our human experience is based in our genetic constitution and its interaction with the environment. They use this to infer that empirical knowledge about our human evolution is profoundly important for moral philosophy because it "renders increasingly less tenable the hypothesis that ethical truths are extrasomatic, in other words divinely placed within the brain or else outside the brain awaiting revelation" (Ruse & Wilson, 1986, p. 174; Kitcher, 2006, p. 577). That is, Wilson and Ruse claim that an empirical understanding of how our ethical concepts evolved is evidence against moral objectivism.

This does not sit well with Kitcher, who points out that an empirical understanding of our ability to make moral judgments should not rule out moral objectivism any more than an empirical understanding of our abilities to make judgments in other areas of inquiry like mathematics, physics and biology should rule out objective truth in these areas (Kitcher, 2006, p. 607). After all, we could trace the history of our ability to understand biology and still believe that there are objective biological facts in the world which we are discovering. It seems the same could be said for our moral understanding as well, unless we view moral inquiry to be fundamentally different from biological inquiry. Now, Kitcher grants that Wilson, Lumsden and Ruse are doing just this, viewing moral inquiry as different from inquiries of mathematics and various sciences, but he notes that they do not support this with an argument. Thus, Kitcher reemphasizes his caution about overinterpretation, stating that any profound consequences of Project 1 endeavors would be due to their metaethical denial of moral objectivism rather than from their empirical gains.

I agree with Kitcher's analysis of this project. On its face, this project emphasizes that science can teach us about how humans developed their moral thinking. We may find that it is connected to different cognitive systems; for example, children's sharing behavior is connected to their counting skills (Chernyak et al., 2018). Findings such as these give us insight into the psychological mechanisms underlying our moral thinking. As Wilson aptly points out, we are no longer dependent on divine explanations to understand how we come about our moral thoughts and inclinations. Those, science has revealed, are the outcomes of natural processes and though we may continue to be astonished at our findings, they no longer carry a supernatural mystery. However, understanding how we come about moral thoughts and inclinations is different than asking whether there are objective moral facts. So, in a response similar to Kitcher, I see this first

project as viable but easily overblown. It is tempting for scientific findings to overreach here because once we understand the "What can science teach us about morality?" question, the very next question is often "Why does it matter?". I believe my approach gives this second question a viable answer – one that does not overinterpret scientific findings about morality but rather shows how these scientific facts, in their own right, are meaningful to us.

PROJECT 2: *Sociobiology can teach us facts about human beings that, in conjunction with moral principles that we already accept, can be used to derive normative principles that we had not yet appreciated* (Kitcher, 2006, p. 576).

In this second project, science gives us facts which can be used with moral principles we already have, and from this we may change our fundamental ethical principles. Kitcher thinks of this project as somewhat uncontroversial but acknowledges that sociobiology is not at work alone in this endeavor. Scientific facts of all sorts can contribute to how we think about moral situations, but Kitcher argues that this does not dismiss the work of philosophers – they are still needed to evaluate competing ethical principles. I agree with Kitcher that philosophers are still needed, and I will increase the pressure he puts on this project's stance by denying that science changes our fundamental ethical principles.

It is not controversial to claim that empirical facts effect how we think about moral situations. Take, for example, the dispute over the Thirty Meter Telescope (TMT) on Hawai‘i Island. Land considered sacred to the indigenous community is set to be used to build an 18-story telescope facility. Many Native Hawaiians have persistently protested the development of the land, resulting in several arrests of elders. Those representing the TMT have invested heavily hoping to continue its development – now halted for over a decade. Central to the arguments of

this case are empirical facts about the ecological and economic impacts on the island, with both sides offering empirical evidence to support their position. However, this is ultimately a dispute between Native Hawaiian cultural and environmental values and Western scientific and technological values. The moral question is often asked: is it permissible to use this contested land for scientific advancement when it will cause pain and suffering to an indigenous community? The use of empirical facts helps each side persuade others in the courts and communities, but inevitably, this dispute boils down to which set of values we place before the other. Kitcher acknowledges this point when he writes, "Yet while amassing answers is a prerequisite for moral decision, there are also issues that apparently have to be resolved by pondering fundamental ethical principles" (Kitcher, 2006, p. 578). Science can help us learn more about the effects our moral decisions may have on the real world, but it cannot tell us what moral decision we should make. As I discuss further in Chapter 2, claiming that empirical facts should have normative consequences commits the naturalistic fallacy. So far, this second project seems to be quite limited in its reach.

Kitcher mentions other types of empirical facts that may be used in our moral decision making. He writes, "It might be suggested that sociobiology has a particularly important contribution to make to this general enterprise, because it can reveal to us our deepest and most entrenched desires" (Kitcher, 2006, p. 578). Here, Kitcher entertains the assertion that science can reveal to us what contributes to human happiness. The idea is that operationalizing desire and studying it scientifically will give us insight into its nature – we'll learn new things about human desire and the happiness that comes from fulfilling it.

As an example of this perspective, we might consider that scientific findings can match our physiological states to our emotional states. Hormones like endorphins, oxytocin, dopamine,

adrenaline, and others have been linked by researchers to positive affective states. If we find that exercise, for example, releases endorphins in the brain, and that endorphins relieve us of pain and lead to feelings of euphoria, and that being free of pain and in a euphoric state means that we're happy, then might we infer that exercise is connected to our happiness?[3] Kitcher is suspicious of this line of reasoning, and I think rightly so. He writes that "the most prominent sociobiological attempts to fathom the springs of human nature are deeply flawed" (Kitcher, 2006, p. 578). He states that to make sense of findings like these, we should approach the empirical findings in an integrated way, bringing together various sciences and social sciences. In the end, Kitcher argues, we will come back to evaluating the values and desires of different people, and for reasons already given, these dilemmas cannot be resolved with scientific facts.

Extending Kitcher's analysis of this second project, I argue that the role of scientific facts in our moral decisions is not to change our moral principles, but to assist us in our deliberative process. There is no microscope that will reveal what we should value. Rather, the (metaphorical) microscope can reveal facts about our psychology and environment that will inform us whether our actions will uphold or betray the values we already hold. So, in contrast to this second project and in addition to Kitcher's view of it, I hold that science cannot tell us anything that will change our normative principles but can give us facts that will help us better uphold them.

PROJECT 3: *Sociobiology can explain what ethics is all about and can settle traditional questions about the objectivity of ethics. In short, sociobiology is the key to metaethics* (Kitcher, 2006, p. 576).

[3] See Mikkelsen et al. (2017) for more about the endorphin hypothesis.

Kitcher views this third project as "deeply confused" (Kitcher, 2006, p. 581). He points out that Wilson oscillates between two inconsistent positions. First, as addressed in the discussion of Project 1, Wilson claims that sociobiology gives us reason to reject moral objectivism, since it frees us from depending on extrasomatic ethical truths. Instead, we can lean on sociobiology to uncover what is really going on in our moral evaluations. This leads us to the limbic system, where we find the deep emotional center of our brain. Wilson writes, "Human emotional responses and the more general ethical practices based on them have been programmed to a substantial degree by natural selection over thousands of generations" (Wilson, 1978, p. 6; Kitcher, 2006, p. 579). Kitcher interprets Wilson's position as an argument for emotivism, in which the content of our ethical statements is exhausted by reformulating them in terms of our emotional reactions (Kitcher, 2006, p. 579).

The second position claims that sociobiological investigation can reveal to us our deepest desires and needs (as discussed in Project 2) as well as how we can correct our short-term moral insights when they lead us astray from them. This suggests that there is something more to our moral claims than a simple reporting of repugnance. Instead, this second position hints at there being deeper beliefs that should serve as a foundation for our moral codes, and this is at odds with the first, simple emotivist, position.

Because Wilson has presented two conflicting positions, Kitcher doubts his assertion that sociobiology will settle traditional metaethical questions. In a charitable reading, however, Kitcher acknowledges that Wilson's arguments put pressure on moral objectivism in a way that aligns with traditional skeptical views. These skeptical views typically point out that if there are objective moral truths, then they must correspond to some moral order which is separate from the natural order. In other words, if we consider "murder is wrong" to be an objective moral truth,

then there should be some moral order which holds that murder actually is wrong. The proposition "murder is wrong" is thus true because it corresponds with this fact of the moral order; if one were to say "murder is right" it would fail to correspond with the moral order and therefore considered false. The skeptic asks what this moral order is – if it's separate from the natural order, how do we have gain knowledge from it? Our epistemic access to this non-natural world seems to require some sort of way to sense or intuit it, and this gets us into mysterious territory.

This skeptical objection is a difficult charge for moral objectivists to answer, but not impossible. Kitcher imagines a few responses moral objectivists may have, basing them on arguments that defend objectivism in mathematics. Mathematical platonism is the position that there is an objective mathematical realm in which mathematical objects exist. According to this view, mathematical objects are abstract, and so – like moral objectivism holds about moral truths – are not part of the natural order.[4] So, when we make mathematical claims, e.g., "2 + 2 = 4," they are true insofar as they correspond with the mathematical realm. Because mathematical objects are considered by platonists to be abstract, platonists also face the challenge of explaining how we can gain knowledge from them. There are many responses from mathematical platonists, ranging from attempts to show how we can gain epistemic access to the mathematical realm even though it's abstract, to claims that we can have mathematical truths without the existence of abstract objects.

Kitcher remarks that we can take an analogous approach in the moral case. Moral objectivists may give accounts of how we can have epistemic access to a non-natural moral order

[4] There are exceptions to this, such as Penelope Maddy's (1990) view that some mathematical objects (i.e., sets) are concrete, or in the natural order and therefore accessible through our perception. Notably, Maddy has since moved away from this view.

or might hold that we can have moral truths without needing them to correspond to a non-natural moral order. Kitcher's point here, which I believe he makes strongly, is that Wilson's skeptical view does not defeat moral objectivism and does not in any way settle traditional questions in metaethics. The question of objectivism remains open, despite the contributions of sociobiology.

I agree with Kitcher's assessment of Wilson's third project. Furthermore, I think that we should be cautious in how we interpret the role of scientific information in our consideration of metaethical questions. When discussing Wilson's claim that sociobiology can reveal our deepest desires, which we should learn to correct our intuitions to meet, Kitcher points out that Wilson's argument "fails to explain what normative standard gives these desires priority or how that standard is grounded in biology" (Kitcher, 2006, p. 580). In other words, Wilson says that sociobiology can show us what we truly desire and how to correct our moral intuitions so that we can (attempt to) meet these desires. However, Kitcher argues, Wilson does not tell us why we should prioritize these deeper desires over our moral intuitions. Wilson provides no normative or biological backing to this claim, which contributes to Kitcher's skepticism that Wilson is resolving traditional metaethical questions.

I will add here that I do not believe it is the role of science to discover our "deepest desires" or anything of the like. Our values and principles are discoverable upon moral reflection and perhaps through a metaphoric microscope, but not a literal one. Instead, what science can do is give us information about our moral thinking and the cognitive capacities that underlie it. It can tell us how we think, not what we value. So in response to Wilson's use of the limbic system, I argue that the role of sociobiology here is to help us understand things like how moral situations may evoke moral judgments with concomitant emotional responses. We may end up caring about things revealed to us through scientific investigation – I might learn about the

limbic system and respond with "Wow! I'm amazed and now I care about that!" However, scientific investigation will not reveal to me that I care about something – I will not learn something and respond with "Wow! Apparently, I care about that – I had no idea!" Given that metaethics is traditionally concerned with questions about what we value, I do not believe that science is the key to metaethics.

PROJECT 4: *Sociobiology can lead us to revise our system of ethical principles, not simply by leading us to accept new derivative statements – as in number 2 above – but by teaching us new fundamental normative principles. In short, sociobiology is not just a source of facts but a source of norms* (Kitcher, 2006, p. 576).

Project 4 is an extension of Project 2, which claims that sociobiology allows us to derive normative principles we didn't already know. Recall that Kitcher considers this project somewhat uncontroversial – he only adds that we should acknowledge that ethicists are still needed to "evaluate the different desires and interests of different people (and, possibly of other organisms)" stating that this quintessentially moral task cannot be discharged by sociobiology (Kitcher, 2006, pp. 578-9). I, however, rejected that Project 2 is uncontroversial because I believe that that the information we gain from science is more limited than Wilson and Kitcher are understanding it to be. I argue that scientific information can be used to help us understand the impact our moral actions will have on the world, but that it is unable to change our foundational ethical principles. So, sociobiology can help us in our moral deliberations, but it will never be able to tell us anything new about our values.

If Project 2 approaches the naturalistic fallacy, as I suggested, then Project 4 is an egregious offender. Kitcher seems to agree with this sentiment, writing that Wilson's writings

fail to give us any reason to think of this project as anything other than a blunder (Kitcher, 2006, 584). The problem Kitcher identifies is that Wilson's work presupposes nonbiological ethics. This is evident when Kitcher summarizes what he views as a fundamental ethical principle proposed by Wilson. That is, Wilson maintains that the scientific fact (S) that the DNA of any individual human being is derived from many people from past generations and will be distributed among many people in future generations gives us reason to accept as a fundamental ethical principle (W) that people should do whatever is required to ensure the survival of the gene pool for Homo sapiens (Kitcher, 2006, 582). This descriptive-to-normative move is a proper example of an is-to-ought violation, and Kitcher subsequently points out places where Wilson has explicitly rejected that such a move would be illicit in the first place.

Kitcher suggests that the only way to save the move from (S) to (W) would be to provide supplemental normative premises to the transition. This, however, would revert to an endeavor under the Project 2 purview, which Kitcher roughly accepts as a legitimate use of sociobiology. Project 4, as it stands, is unacceptable.

To make matters worse for Project 4, Kitcher points out that the principle (W) invites – or perhaps even demands – actions we consider morally suspect. The idea that people should do "whatever is required to ensure the survival of the gene pool" is hardly a foundational ethical principle we can rally behind; it welcomes rape and coercion. It assumes that ensuring the survival of the gene pool should be valued more than our own bodily autonomy or reproductive autonomy. Thus, in addition to the shocking inaptness of (W), the nonbiological ethical judgment that it should be privileged above other values shows that this view is self-defeating. We would still need ethicists to determine which values should be placed above others in instances of conflict.

Ultimately, if (W) is an example of the sort of normative values that we can derive from scientific facts, then we are headed in a precarious direction. We would end up with at least some "values" that violate our moral sensibility. Appealing to their natural status to privilege them over the ethical principles we are committed to, takes the naturalistic fallacy to a new level. Not only is Wilson claiming that we can derive an ought from an is but he's deriving an ought that violates our other oughts. Science is not just improving moral codes at this point, as the earlier projects suggest. By deriving foundational ethical principles that feel wrong to us, it is changing the very nature of moral codes and the role these codes play in our lives.

So where do these four projects leave Kitcher's assessment of how to biologicize ethics? Project 1 explains how people acquire ethical concepts, which is fairly uncontroversial, but we need to remember to keep our conclusions tempered. Project 2 allows us to derive from scientific findings normative principles we didn't already know. Kitcher found this to be somewhat uncontroversial though he has doubts that sociobiology can do this without leaning on ethicists. Additionally, I disagree that Project 2 is permissible in any sense as it edges on an illicit jump from is to ought. Project 3 claims that sociobiology can resolve traditional metaethical questions, and Kitcher found this to be inaccurate and deeply confused; I agree. Project 4 claims that sociobiology can teach us new fundamental principles. Kitcher strongly rejects this notion, as do I.

Kitcher's advice to those who want to contribute to the sociobiological work on ethics is that they should be clear about what sort of project they are undertaking and respond responsibly. Those whose work aligns with Project 1 should be reflective of the methods they use to construct histories of human moral thinking, tempering their conclusions and learning from past blunders (e.g., neo-Darwinists). Those whose work aligns with Project 2 should "explicitly acknowledge

the need to draw on extrabiological moral principles" as well as reflect on whether the question they're asking is best answered by sociobiology (Kitcher, 2006, p. 585). Those concerned with Project 3 need to reflect on their metaethical stance, asking themselves if they believe that moral statements can be true or false. If they do believe moral statements can be true or false, then they should be prepared to explain how it is grounded. If they do not, then they should be prepared to explain what moral statements are and how we determine which ones should be privileged. Finally, those who undertake Project 4 will need to present a solid case for why the move from the is of biology to the ought of morality can be justified in any sense. It is not enough to reject the validity of the naturalistic fallacy; a persuasive case must be made here if anyone is to accept that biology can be the source of foundational ethical principles.

As I presented Kitcher's assessment of the biologicizing of ethics, I added doubt to several aspects of these projects. My overall response to Kitcher's chapter is that each project stops short of providing us with a meaningful account of how science can contribute to ethics. I believe that Project 1, in which sociobiology aims to explain how people acquire ethical concepts, is the most legitimate endeavor, as long as we heed Kitcher's warning not to overinterpret the claims of its descriptive findings. This may feel unsatisfying, especially when this endeavor is compared to its provocative successors. So, I now present a *new* project – one that picks up where Project 1 leaves off.

A NEW PROJECT: *Sociobiology can teach us about how our cognitive processes cause us to unintentionally drift away from our deeply held moral principles. This inquiry allows us to make practical changes to increase our agency over our moral behaviors.*

My proposed project adds a claim of significance for the scientific understanding of how we acquire ethical concepts, which was missing from Kitcher's project. The information we learn about the evolution of our ethical concepts is important not because it will tell us anything about the metaethical status of ethical truths, or unveil our deepest desires, but because of its practical value: it will contribute to our moral deliberations in concrete ways.

Sociobiology, and specifically moral cognition, are ripe with fascinating and surprising findings about the way humans think about morality. This was not lost on Kitcher, but his responses to the outlined projects fall short of seeing these findings to their full potential. The value in our scientific research on morality is *practical*. Understanding how we make moral judgments, how powerful our moral beliefs are, and how we form moral intuitions, gives us the opportunity to enhance our moral deliberations. We will have a fighting chance to *at least sometimes* correct for unconscious processes that pull our moral behaviors away from our moral principles.

1.2 Concrete Changes

To show this new project in action, I offer concrete examples in Chapters 3 and 4. Through these cases, I argue that knowledge of scientific facts about our moral cognition can and should influence our moral deliberations. Importantly, insofar as these cases support the new project and not Project 1, I argue that the science tells us about our moral cognitive processes – not about morality itself. So, whereas those supporting Project 1 tend to inflate the significance of scientific information about our moral concepts by claiming that it can be used to tell us about morality itself (i.e., refuting moral objectivism), I find that the empirical picture is significant

because we can do something important with it. I argue that we should use this information when we deliberate about moral situations. It might occur to the reader at this point that I have been critical of other writers' usage of shoulds and oughts. So to be clear, the ought in my argument comes from appealing to how this scientific knowledge can help us change our moral deliberations in ways that will make us more satisfied in the long-term. In other words, I am not arguing that science tells us we should listen to science. Rather, I am aiming to persuade the reader that implementing the knowledge she gains will contribute to her overall satisfaction. This appeal is philosophical, not biological.

Two cases I present show that our actual moral behaviors can dissociate from our moral theories and principles. That is, we think certain moral values are important to us and that they guide our behaviors, but empirical investigations show that this is not always how it works. Instead, our moral behaviors have many influences of which we are unaware. By understanding how our moral thinking and behavior has been shaped by our evolutionary history and continues to be influenced by our social environment, we begin to unveil the innerworkings of our moral cognition. In this respect, science is illuminating our understanding of our moral processes, which this section has shown is a fairly uncontroversial if not mundane project. The significance of this information lies in what we do with it. Below is a preview of the cases I highlight to show how individuals can utilize these empirical findings in their moral deliberations, leading them to make concrete changes with which they will be more satisfied with upon reflection.

My first case examines the moral regret that comes with believing one has made a "wrong" choice in a moral dilemma. Moral dilemmas, e.g., the Trolley Problem, are commonly used by moral philosophers to frame their views about normative theories. These dilemmas are presented as difficult problems to solve, where two competing moral theories are pitted against

each other. If we just reason and reflect enough, the story goes, we can figure out the right answer. Here, I highlight cognitive neuroscience data that informs Joshua Greene's (2014) dual-process theory, which claims that we have two evolved systems that pull us in different directions in some difficult moral situations. I argue that instead of thinking of moral dilemmas as problems with a single right answer that we can reach through careful thinking, we should recognize that the difficulty arises insofar as our two evolved systems are giving us different answers to the problem. In these dilemmas, we are forced to choose one action over another, resulting in us satisfying one system over the other. The moral regret we feel, I argue, is caused by the remnants of the unsatisfied system. We feel torn because we were being pulled in two different directions. Choosing one over the other does not erase the longing toward a path we did not follow.

In my examination of moral regret, I reframe what our regret points to. In the example I use, I explain that the person plagued with regret was in such a state because he thought of his regret as evidence that he made the wrong choice. My reframing points out that he would have felt regret with either of the two actions he could have made – so it is not useful to think of his regret in this way. Instead, we should think of his regret as the longing of his unsatisfied system for the consequences it promised. Furthermore, incorporating this perspective into our moral deliberations can help us to consider the consequences of future dilemmas. That is, when we reflect on our moral activities, we can better consider how we might feel if we were left with a seemingly impossible choice of choosing to satisfy one system and resist the other. We may consider our principles and how they align with the feeling of regret over the dismissal of one system's recommended actions.

The second case directs us to another surprising tendency: we tend to revise our factual beliefs to better support our moral judgments. The intuitive position is that we revise our *moral* judgments to better align with our *factual* beliefs. Consider, for example, two people having a dispute about mandatory minimum sentences. One cites factual data about recidivism rates to support their claim that mandatory minimums deter potential offenders from committing crimes. The other cites factual data about the rate of non-violent offenders being given extensive prison terms to support their claim that mandatory minimums punish crimes unfairly. In both cases, the parties are presenting their factual beliefs as support for their moral positions about the use of mandatory minimum sentences. The regular use of factual data as support for arguments intended to persuade others indicates people often believe that a change in factual belief should cause a change in moral judgment.

However, research by Thomas, Stanford and Sarnecka (2016) and Liu and Ditto (2012) challenges the intuitive position. That is, when a conflict arises between one's own factual beliefs and moral judgments, people often change their factual beliefs to align them with their moral judgments. Consequently, we may justify our moral judgment by pointing to our factual beliefs, but in reality, those factual beliefs do not represent a simple descriptive understanding of the world. Instead, Thomas et al. find that there is a tendency for people to revise their factual beliefs so that they better support their own convictions (Thomas et al., 2016, p. 13). I argue that being aware of this tendency should push us to consider the balance between our moral and factual beliefs in our moral deliberations. If we use factual data in our moral deliberations, then we should deeply scrutinize whether these factual beliefs describe the world objectively. Not doing so would undermine the role we think they have in supporting our moral judgments. But once we begin to look more closely at this process, we find that these factual beliefs have likely

been filtered through our moral judgments. If we believe that the causal direction should flow from factual beliefs to moral judgment and not vice versa, then we should be skeptical about invoking facts when seeking to support our moral judgments.

The third case examines the power of intuitions in our moral judgments. Here, I look at two tendencies regarding group membership: our tendency to favor ingroup members and disfavor outgroup members, as well as our tendency to punish those who fail to conform their respective group norms. In both tendencies, group membership is constructed under totally minimal conditions – even something as simple as assigning someone to a group by giving them a sticker can elicit these ingroup/outgroup responses. Because we may consider such minimal grouping to be morally irrelevant, we should be skeptical of our own moral judgments in situations where these tendencies may be evoked. Before previewing my third case, I briefly describe these tendencies and how our skepticism can and should affect our moral deliberations. The tendency to favor members of our ingroup and disfavor members of our outgroups, even under totally minimal conditions, means that group membership can play a role in our moral judgments, even when we think the grouping is morally irrelevant. I argue that we should use this knowledge in our moral deliberations, being skeptical of our own moral judgments toward outgroup members whose membership status is not morally relevant. That is, if I have a moral judgment toward an outgroup member who is different from me in a way that I find morally irrelevant, e.g., we are of different races, then I should be more careful in deliberating about the moral judgment. I should reflect on the fact that I am more likely to disfavor this person for reasons that are irrelevant to the moral judgment, and soften or alter the judgment accordingly. By understanding how this mechanism of preference and bias works, we can a better align our

moral judgments with our moral principles: we will increase fidelity to our moral commitments by decreasing fidelity to our moral intuitions.

The tendency we have to punish those who fail to conform to their respective group norms emerges in the prescriptions we build out of our descriptive observations of groups. Recent work in cognitive development shows that when children observe group regularities, they form negative evaluations of group members who do not conform to the regularity. In other words, children do not just expect group members to act similarly, they think they should act similarly. This tendency has been found to be quite robust. Though it declines with age, it is still often present in adults as well. I argue that this descriptive-to-prescriptive tendency presents another opportunity to pause and reflect. If we find that the slide from descriptive-to-prescriptive is morally irrelevant, then we should resist it. For example, if I find myself judging a woman negatively for being assertive, I should pause and consider that I probably would not judge a man negatively, or as negatively, for being assertive. The descriptive-to-prescriptive tendency illuminates this for me: I have observed men to be assertive much more often than I have observed women to be assertive, and these observations have turned into prescriptions. In my moral deliberation, I should recognize that my negative judgment is probably being emphasized by the woman not conforming to how I think women should act. I do not think there is a moral difference between men and women, so I should not find the behavior of the same act as worse in a woman than in a man. So, in a similar spirit as the ingroup/outgroup bias tendency, my moral deliberation is being influenced by factors that I, myself, reject as morally irrelevant. Understanding the tendency to punish those who do not conform to their respective group norms will allow me to better align my moral judgments with my moral principles. Understanding these tendencies shows us that our moral intuitions can well up, pointing us in various directions –

some of which are in tension with our reflectively-endorsed and more stable moral commitments. Understanding how these moral intuitions have evolved helps us to sift through them so that we may counteract those intuitions we find to be contrary to our commitments. This allows us to make moral decisions that are more in line with our moral values.

1.3 Christine Korsgaard's *Normative Self-Government*

> And it is not a small difference, that ability to be motivated by an ought… A form of life governed by principles and values is a very different thing from a form of life governed by instinct, desire, and emotion – even a very intelligent and sociable form of life governed by instinct, desire, and emotion.
>
> - Christine Korsgaard
> in *Primates and Philosophers*, 2006

Korsgaard's approach to morality differs from those we have examined thus far. Rather than focusing on the ontological status of morality (i.e., whether there is a moral reality that exists in a theological or natural order), Korsgaard views morality as a human practice grounded in reason. Because she views it as a human practice, Korsgaard has been prompted to discern whether morality is rooted in our past or whether it represents a break with our past (Korsgaard, 2006, p. 99). That is, in a chapter of Frans de Waal's book *Primates and Philosophers*, Korsgaard responds to the question of whether morality is a uniquely human capacity or if it is shared, even gradually, with other animals. In response, she examines the essence of morality, which brings her to our human capacity for normative self-government, which she has since elaborated on in her book *Self-Constitution: Agency, Identity and Integrity*. In this section, I introduce this capacity and how it relates to morality. In doing so, I find that by implementing scientific knowledge about our moral cognition into our moral deliberations, we will enlighten and empower our capacity to self-govern. Thus, I consider my view as a friendly amendment to

Korsgaard's; if self-government lies at the essence of morality, then knowing how to better self-legislate – which I believe my view accomplishes – seems a clear virtue.

Korsgaard develops her picture of normative self-government from the Kantian view of autonomy. For Kant, we are autonomous when we make our own laws. For example, consider an individual who knows that an action A would produce a pleasant event E. If the individual is heteronomous, then he is guided by laws that are outside himself. He may be guided by a law of pleasure, which tells him he must do A because it will lead him to E, which is a pleasurable event. An autonomous individual, however, is only guided by laws she has made for herself. So she will not be governed by a law of pleasure unless she has chosen to be so governed. When the autonomous individual acts, she is acting on a law she has legislated for herself – she self-governs, with nothing outside her giving her any laws (Korsgaard, 2009, p. 153).

Our capacity for normative self-governance allows us to not only have intentions, Korsgaard says, but to assess and adopt them. So, through our autonomy, we choose the laws we follow and this is where morality emerges. She writes, "The morality of your action is not a function of the content of your intentions. It is a function of the exercise of normative self-government" (Korsgaard, 2006, p. 112). So the morality of our actions is not reflective of whether our intentions are good or bad. Rather, it is reflective of the autonomy we employ by choosing what laws we allow to govern us. A dog, for example, may consciously and intelligently pursue ends that were given to him by his affective states – he may work to open a bag of dog food to satisfy his hunger. Humans, though, pursue our ends at a deeper level. We determine whether we should make hunger a law that governs us, and that is where the morality comes in. Morality emerges not through what we intend to do, but rather through what we legislate for ourselves.

When contemplating the question about whether nonhuman animals may possibly exhibit morality in this sense, Korsgaard gives a tempered reply. She writes, “There is nothing unnatural, nonnatural, or mystical about the capacity for normative self-government. What it requires is a certain form of self-consciousness: namely, consciousness of the grounds on which you propose to act as grounds. (Korsgaard, 2006, p. 113). By this, Korsgaard means to point out that a nonhuman animal, our dog for example, may be conscious of the food that he desires, and that it is something he is trying to eat. But the dog does not seem to be conscious that he desires the food and that he may destroy the dogfood bag in order to eat it. The dog, as far as we know, does not contemplate, “Well, should I go eat that food? I know I’m going to destroy this bag to get it, but does wanting to eat it so badly really give me reason to destroy the bag?” This consciousness of the grounds as grounds is what powers our normative self-government. If I see a cheesecake, I will almost certainly ask myself whether my affective state gives me reason to eat it. I will contemplate, “Well, should I eat that cheesecake? I’m inclined to eat it, but is that reason enough to actually eat it?” Korsgaard says that at this point, we are in a position to raise a normative question about what we ought to do (Korsgaard, 2006, p. 113). At this point, I will reason about whether I ought to eat the cheesecake. I may have competing values or ends – I may have promised my partner that I would wait for them before I ate it. In reasoning, I will look inward, focusing on whether my potential action is justified by my motives or whether my inferences are justified by my beliefs – did my partner actually say they want me to wait, or am I inferring that based on past experiences? This gives Korsgaard reason to believe that the capacity for normative self-government is something unique to humans, which implies that morality is unique to humans.

The uniqueness question is interesting here, but only indirectly so. For this view I propose, the question of whether nonhuman animals exhibit moral behavior or have a proto-morality is not relevant – I do not have any skin in that game. However, insofar as this prompts Korsgaard's search for what exactly constitutes morality, it is crucial. She writes about the capacity for normative self-government, "And it is in the proper use of this capacity – the ability to form and act on judgments of what we ought to do – that the essence of morality lies, not in altruism or the pursuit of the greater good" (Korsgaard, 2006, p. 116). For Korsgaard, morality is about our ability to form and act on judgments of what we ought to do. I believe that my view will only empower this ability.

Understanding our moral cognitive processes will allow us to improve our self-government by giving us a better-informed choice of laws we may allow to govern us. To illustrate the influence my friendly amendment has, I begin with the experience of moral behavior, from intuition to reflection, through Korsgaard's picture. I then show the same path through my own view, and I argue that my picture gives us benefits we cannot ignore.
Under Korsgaard's picture, we begin with an inclination to act in a certain way. We choose the laws that we will allow to govern us, we reason about whether we have the grounds for the action we are considering, and we form judgments about what we ought to do. For example, say that I am an employer who wishes to hire one of the ten people I just interviewed. Through my autonomy and self-government, I have chosen fairness as a law to be governed by. Setting aside that there may be legal requirements from my political state about hiring practices, I have made a principled decision that fairness is important to me and that I should be open to any candidate based on their qualifications alone and not any matters of personal identity. I reason about whether I have the grounds for hiring the candidate I liked most. She has terrific qualifications,

passed the background check, and had a personality that will likely fit our work environment. So, I form a judgment about what I ought to do. I may ask myself whether I gave the other candidates a fair chance, or whether I should broaden my search in case a better suited candidate shows up in the next round of interviews.

Assuming I follow through with the action and I hire the candidate I liked most, my reflection later on will likely go one of two ways: either I find that it worked out well or I think of it as a mistake. If it worked out well, I will probably not think much more of it, and may use it during my reasoning process in making a similar decision in the future. If it doesn't work out well, say the employee steals from the company, I will most likely regret my choice. I will look back on my decision and ask myself where I went wrong. I will wonder how to avoid making a similar decision in the future – should I not hire someone who I like in the interview? The process of making a decision and later reflecting on it in Korsgaard's picture is fine. It allows for us to use our reason to decide what to do, and this opens our deliberative process up to examine our own motives and beliefs.

What I would like to add to Korsgaard's model is our incorporation of scientific information about our moral cognitive processes. Going through the same scenario highlights the contribution that my view can add and the benefits it will induce. It begins the same: I am an employer looking to add to my team of employees. I am committed to fair hiring practices, so I allow fairness to govern my deliberative process. However, I am also aware of research that shows people have biases in favor of members of their own group and against members of other groups. Even in minimal conditions, where classes like race, gender and religion are controlled for, people tend to favor those in their own group over those outside their group – even if they

know that their in-group member will act badly![5] So, when I begin to reason about whether I have grounds to hire the candidate I liked, my deliberation will now include several more aspects than it did in the previous scenario. Sure, her qualifications are good, she passed the background check and her personality fits the office. But am I favoring her over other candidates because I see her as a part of my group and the others outside it? Is she similar in gender, race, ethnicity, ability-status, age, or any other protected-classes? In my previous deliberation, I may have noticed such things, but without understanding how deeply in-group favoring and out-group favoring go, I may have been more likely to dismiss this as evidence. Knowing how deeply these tendencies reach, then, I should be even more inquisitive of my intentions. Did she go to the same college as me? Is she my height, or does she have the same hairstyle? Now, when I form my judgment about what I ought to do, my deliberative process is highly informed. I may give more weight to my concern that I didn't give the other candidates a fair enough chance. I may think about the selection process for who was invited for interviews and decide that it should be broadened to avoid in-group biases I believe to be unfair.

There are two important issues here that I want to address before continuing into the reflection stage of my moral deliberation. First, it is crucial to note that the scientific information I am using is related to the law I have chosen myself to be governed by. This was an autonomous decision, and it is not the science that is telling me what laws I should be governed by. The naturalistic fallacy, therefore, is not at work here. Secondly, when I am deciding what counts as grounds here, I am deciding that what I believe to be morally irrelevant factors are impeding my moral decision. That is, I do not think that it's morally relevant whether a job candidate went to

[5] Here, I am alluding to a study by Baron and Dunham (2015) which I discuss in more detail in Chapter 4.

the same college as me. So, I find that my commitment to fairness is being confounded by a morally irrelevant factor. This is different from weighing values against one another, as happens in the next stage, where I decide what I ought to do. Thus, my picture changes our understanding of the evidence itself, which has the effect of changing our deliberative process. However, the scientific information is not affecting what I value. I am still choosing to be governed by fairness and I am still weighing oughts against one another. The difference here is I have a greater understanding of the cognitive processes that power my intentions, or inform what I consider grounds for my desire to act.

The information I learn about my cognitive processes can make an impact on my reflection as well. In Korsgaard's picture, I may be moderately reflective, especially if the consequence of my decision was unwanted. In my picture, however, we can come to understand our past decisions in light of what we continue to learn about our moral cognition. Consider the scenario where I hired the candidate I liked, only to find that she stole hundreds of thousands of dollars from the company. I may blame myself for hiring her, for trusting her. I may feel guilt and regret for bringing her into my workplace and around my other employees, who trusted me to make a good hiring decision. Incorporating scientific knowledge about in-group/out-group biases into my reflective process will help ameliorate the guilt and regret I feel. By understanding the factors involved in my preference of this candidate over others, I can understand how it clouded my deliberative process. These were not factors I was aware of, and they're not even issues I consider to be morally relevant. Without the scientific understanding of how my moral cognition works, I would have been ignorant to their involvement in my deliberative process altogether. Now, though, I can see my decision as one that was affected by factors that I understandably did not foresee.

Two questions arise in our consideration. First, is not ignorance bliss? Would I not have been happier had I not understood the way in which my actions were biased? Without understanding how my biases affected my judgment, I could have continued forth, incorporating biases I am completely unaware of into decisions I think I am carefully wading through. The science complicates this. It calls me out, not by telling me what laws should govern me or by saying I am wrong to choose one ought over another. Instead, it tells me that my deliberative process doesn't work the way I think it does. The input I am feeding my deliberative process is faulty, and this is going to change the outcomes of my behaviors. I do not realize that I am doing this, but I am doing it nonetheless. If I put faulty information into my deliberative process, then my judgment process will not actually be aimed in the direction I am looking and I will misfire. Now, this is not to say that I will have perfect aim with good, or well-informed, input. But I will have a better chance at actually achieving what I set out to do. This may make things more difficult and will certainly reduce blissful ignorance. However, I will improve my moral deliberative process, strengthening my ability to self-legislate.

The second question we may ask is whether appealing to descriptive facts is actually a way to evade moral culpability. Imagine here that I am confronted by someone who has lost their job as a result of the employee's theft. This person may blame me, asking why I hired the employee over the other job candidates. If I respond by explaining that she went to my same college, and I didn't realize at the time how that might bias me, but now I know more about moral cognition so I will be better prepared next time, this will be of little consolation to the now-unemployed person. In fact, they may accuse me of trying to justify my choice or evade responsibility for my role in her hiring altogether. "You can't just blame this on science!" I can hear them say. And this is a fair point. Backlash can occur when scientific explanations are given

for deviant behavior.[6] My response here is that we need to maintain a careful boundary between the descriptive and normative parts of this process. When an appeal to scientific evidence about our moral cognition is used to justify or excuse an action, this is a violation of that boundary. The problem was with the faulty input data, not with the judgment system. But that faulty input data has real-world consequences insofar as it informs our judgment process, which guide our actions. So admitting that there was a problem in the deliberative process and working to fix it seems to be a responsible thing to do. The alternative is to fixate on the judgment process without acknowledging the role that the absence of descriptive information played. I believe we should reflect on our judgment process and the different oughts that we weigh. However, without strengthening our understanding of how our moral cognition works, we will always be working with faulty input, and so our deliberative process will be underpowered.

In this section, I aimed to show how Korsgaard's notion of our capacity for normative self-governance can be improved by looking to science to help us understand our moral cognition better. By inviting scientific information about our moral processes into our moral deliberative process, we will strengthen our ability to self-legislate. We will understand how our moral processes are sometimes affected by factors we do not suspect would have any influence. Importantly, knowing about these factors can help us correct for them when we are deliberating an action. We should be careful, though, how we use the scientific information in this process. The role of the scientific information is to illuminate how our moral processes work, not to tell us which laws should govern our actions or which values should be weighed more heavily than

[6] For example, there has been criticism over the scientific study of pedophilia. Research from neuropsychology has shown differences in brain regions between pedophilic and nonpedophilic men (Cantor et al., 2007). These studies have provoked controversy among the public (see Sapolsky, 2017).

others. I believe that this use of scientific information has clear benefits, as it can help us to better understand ourselves and to make decisions that better exhibit our values.

CHAPTER 2: Illicit Border Crossings

2.0 What Science Can't Teach Us About Morals

> In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. (Hume, 1739/2000, T3.1.2.27)
>
> -David Hume
> *A Treatise on Human Nature* 1739

When considering the role that scientific knowledge can play in our moral thinking, it is important that we closely examine the limits that curtail the conclusions we may fairly draw. David Hume famously remarks that in every normative account of ethics he has encountered, the author shifts from making a case for what is or is not to making a case for what ought or ought not to be. Hume's puzzlement has grown in its influence, becoming known in various iterations as the is-ought problem, Hume's Law and the naturalistic fallacy. Ultimately, Hume's remarks have become a warning to philosophers, requiring them to base any normative ethical claim on more than simply descriptive evidence.

I consider Hume to have drawn a line in the sand, clearly demarcating the limits of descriptive power in moral philosophy. In other words, I think Hume was right to question normative claims that depend solely on descriptive evidence. To do so would be to conflate what is good with what is natural. In this chapter, I examine two naturalistic theories of morality which depend on an empirical understanding of ourselves and our environment to make moral claims. This discussion gets dangerously close to Hume's line in the sand, and I argue that these two accounts overstep that line. I use these accounts as cautionary tales, ensuring that my

argument does not follow in their illicit move. Instead, my position firmly holds that our scientific understanding of moral cognition cannot tell us how we ought to act.

2.1 Philip Kitcher's *Pragmatic Naturalism*

In his 2011 book, Philip Kitcher uses a naturalistic framework to introduce the reader to the ethical project. This project, Kitcher argues, has emerged from the human social situation, evolving over a period of tens of thousands of years. In learning to live together, we have developed ethical practices in response to needs and desires that arise as a part of our human existence (Kitcher, 2011, p. 8). Kitcher constructs a how possibly narrative, giving descriptive evidence of how the ethical project emerged. Then, through some updates to metaethical notions, he concludes that the ethical project leads us to consequentialism. In this section, I outline Kitcher's project and then present an argument from O'Connor et al. (2012) that makes the case that Kitcher's theory commits the naturalistic fallacy. I agree with O'Connor et al. on this point and extend their reasoning in the following section 2.2, to apply to Joshua Greene's dual-process theory of moral judgment.

Desiring to revolutionize the ethical narrative, Kitcher introduces a new way of thinking about ethics: pragmatic naturalism. His theory begins in a similar way as other naturalistic accounts of ethics: seeking to explain moral cognition in terms of evolution. Such theories offer descriptive accounts of ethics, treating moral facts as evolutionary facts. While naturalist accounts overlap in much of their descriptive story, they often differ in terms of their prescriptive scope. Kitcher's theory begins as an ethical project, grown from the human social situation

(Kitcher, 2011, p. 3). We will find that the ethical project turns into a prescription for a particular normative account: consequentialism.

Kitcher views our ethical project as emerging with the development of altruism in humans. This is because altruism, or the ability to identify the needs of conspecifics and act in a way that will benefit them, is advantageous to individuals and is ultimately rewarded by the group. In human and non-human animal groups alike, individuals inevitably fail to respond altruistically to a conspecific, resulting in an altruism failure. Whereas non-human animals require costly acts like many hours of grooming to repair relations after an altruism failure, humans have developed an especially efficient mechanism for avoiding altruism failures: normative guidance. Normative guidance tells group members how to act so that they can be seen by conspecifics as behaving altruistically (Kitcher, 2011, p. 74). Because of its effectiveness in diminishing altruism failures, Kitcher credits normative guidance with allowing human groups to build social relationships unavailable to our primate cousins.

Once normative guidance began to take hold, members of social groups began to explicitly agree upon codes of behavior. According to Kitcher, egalitarianism was central to the formulation of these codes because the altruistic foundation of normative guidance required that group members show a vested interest in the well-being of other group members. Preventing the contributions of fellow group members to the ethical discussion would be antithetical to the altruistic behavior endorsed by the group. So, Kitcher's story of the ethical project seems well on its way, with one very difficult challenge.

Although normative guidance can explain how human behavior changed in a way that welcomed the ethical project, it cannot explain the private thoughts or subconscious underpinnings of the behavior. Human motivation is especially important in our understanding of

ethics; we think of ourselves as being ethically driven by our inner selves, not by our fear of abandonment by our group. So how, then, can Kitcher explain the emergence of ethics if all normative guidance can do is point to behavior modification? To answer this, Kitcher restructures how we think about ethics, challenging the metaethical notions of ethical truth and ethical progress.

The conventional idea of ethical progress is that it is an accumulation of ethical truth: we discover ethical truths, and in doing so make progress toward some ethical end. Kitcher's metaethical restructure reverses these roles. He proposes that "ethical progress is prior to ethical truth, and truth is what you get by making progressive steps" (Kitcher, 2011, p. 210). We begin with the notion of ethical progress, which is best seen as a refinement stemming from the original function of the ethical project: preventing altruism failures. As altruism failures become less frequent, social harmony increases. The ethical codes that have arisen to keep us in line with the original function can now become refined and lead to new, derived functions. Social technology emerges, where derived functions develop to resolve the needs created by the previous functions, like a seatbelt serving a function that did not exist before moving vehicles.

One may object at this point, noting that all Kitcher has given us so far is mere change and not progress. If this change is not connected to a notion of ethical truth, we are unable to distinguish between "good" and "bad" changes. The term progress surely implies a good change; after all, we would not call slavery ethical progress, but we certainly reserve the term for its abolishment. So, what happens when there is a conflict between derived functions? How do we choose one over the other? For that, we need a normative account of the ethical project – a way to justify certain changes as progressive and others as regressive.

Before his normative account can get off the ground, Kitcher runs up against a class of objections that thus far has been insurmountable for naturalists: the naturalistic fallacy.[7] This fallacy traces its roots to an objection put forward by Hume, viz. that it is unjustified to derive "ought statements" from "is statements" (Kitcher, 2011, p. 254). In other words, naturalists can give a straightforward descriptive picture of how ethics evolved, but any move to prescribe actions – any push for a normative theory – cannot be justified by that descriptive account alone.

Kitcher takes great care in examining different versions of the naturalistic fallacy, and in the end, he claims that the restructured ethical framework of pragmatic naturalism allows it to avoid this Humean challenge altogether. Recall that traditionally, ethical progress has been viewed as an advancement toward ethical truth – truth being the fundamental notion. So, in the traditional account, a descriptive "is" story leads to a prescriptive "ought" story in which the prescriptions are in line with ideas of ethical truth, or how things "should" be. This inference is problematic because a normative conclusion is drawn from descriptive evidence; there is no normative evidence provided to substantiate a normative conclusion. However, in reversing the roles of progress and truth, Kitcher presents an account in which a descriptive "is" story leads to a prescriptive "ought" story where the prescriptions are in line with ideas of ethical progress, or how to respond to the original and derived functions. Kitcher's revised framework has reformulated the problem, avoiding any Humean mystery. As Kitcher remarks, "Once ethics is viewed as a social technology, directed at particular functions, recognizable facts about how those functions can better be served can be adduced in inferences justifying ethical novelties" (Kitcher, 2011, p. 262). It does not seem mysterious to say that we should use seatbelts based on

[7] Kitcher notes that there is not one "naturalistic fallacy" but rather various mistakes to which this name is attached (Kitcher, 2011, 253). I am using the term here as a *category* of fallacies sharing the "is/ought" problem.

a factual assessment of modern transportation. Similarly, the prescriptions of pragmatic naturalism are simply promoting what will best serve the functions of the descriptive story of ethics.

We have now heard Kitcher's descriptive account of ethics and his story about why a prescriptive move is both necessary and justified; but what exactly is his prescription? In articulating his normative stance, Kitcher is clear to separate the normative ethics of pragmatic naturalism from that of traditional accounts. Traditionally, normative ethics has aimed to "offer a set of resources to help people live as they should" (Kitcher, 2011, p. 285). Both religious and philosophical traditions have provided normative accounts that guide actions and divide good ways to live from bad. Though they differ in the type of authority they invoke, both traditions appeal to external sources of ethical authority. They share a "static vision" in which "correct principles and precepts await discovery, and once apprehended they can be graven in stone" (Kitcher, 2011, p. 285). Pragmatic naturalism, on the other hand, looks only inward, at human evolution to answer questions about the nature of ethics. There is no authority; no static vision of how we should live. Rather, the ethical project constantly evolves; its progress secured by our responses to new functions that have emerged.

In looking for a normative stance that will encapsulate the progress of the ethical project, Kitcher returns to the technology analogy. It would be strange to say that technological progress is measured in relation to some fixed goal; rather it evolves constantly, and progress is attained when the problems of previous technological functions are solved. Similarly, the ethical project evolves constantly, and progress is attained when the problems of previous ethical functions are solved. Any normative account for pragmatic naturalism should be flexible enough to evolve along with the ethical project. Kitcher finds his normative stance in consequentialism.

Citing J.S. Mill, Kitcher defines consequentialism as the belief that "the rightness of actions depends on their consequences" (Kitcher, 2011, p. 289). He argues that it is unjustified to follow deontological ethical systems that push obedience to prior rules without considering the consequences of actions. The responsible thing to do is to follow rules that have been recognized as "well adapted to producing good outcomes" (Kitcher, 2011, p. 289). Recall Kitcher's concept of normative guidance: altruistic behaviors are reinforced by normative guidance, which allows us to explain the evolution of the ethical project without appealing to the "psychological myth" of the "ethical point of view." Altruism may become internally motivated, but this is not the focus of Kitcher's ethical theory. Rather, his focus is on behaviors and actions; consequentialism is a natural fit for pragmatic naturalism.

It is also important to pragmatic naturalism that it adopt an ethical theory that can evolve alongside ethical progress and truth. So, Kitcher specifically endorses a dynamic consequentialism. He writes:

> [A] consequentialist ethical theory can explicitly acknowledge it has no complete specification of the good, seeing its judgments as incomplete and provisional. Dynamic consequentialism makes exactly that admission, supposing that concepts of the good evolve, that some of the transitions among those conceptions are progressive… and that later conceptions of the good are (sometimes) superior to their predecessors, even though none can claim to be the last word. (Kitcher, 2011, pp. 288-9)

Here we see an ethical theory that does not adhere to an external, static notion of truth. Instead, it is flexible and engaged with our notions of ethical progress and truth. Dynamic consequentialism prescribes actions that will advance the ethical project. In doing so, it is aligned with his defense of pragmatic naturalism against the Humean challenge. Dynamic consequentialism does not pull "oughts" from notions of ethical truth, or some picture of how things "should" be. Instead, its

"oughts" tell us how to advance ethical progress; how to better serve the original and derived functions.

Now that Kitcher has made his case for how the ethical project evolved, why it leads us to dynamic consequentialism and why his normative stance avoids the naturalistic fallacy, he leaves us with one more prescription: we must renew the project. We are thousands of years removed from the way our ancestors first responded to the original function. Derived functions grow exponentially, and our response to conflicts between functions has led us to progress in myriad ways. To continue the work of the ethical project, Kitcher calls us to gather once more, continuing the conversation of our ancestors. Because there is no ethical authority, it is imperative that we all participate in ethical discussions of how to best serve the original and derived functions. To give us a nudge, Kitcher offers some preliminary suggestions of how we might continue the ethical project. Most importantly, he recalls the centrality of egalitarianism to the ethical discussions of our ancestors. Kitcher reasons that renewing the ethical project will mean distributing resources in a way that allows everyone a "serious, and approximately equal, opportunity for a worthwhile life" (Kitcher, 2011, p. 396). With this spirit of egalitarianism, we can continue, as a species, to progress in our ethical life.

Deus Ex Machina

In their paper, O'Connor et al. question whether Kitcher's metaethical project successfully connects his genealogical account with his normative stance. Without this metaethical glue, Kitcher's defense against the naturalistic fallacy is powerless, and his account of the ethical project is no stronger than other naturalistic theories. In the most powerful line of their criticism, O'Connor et al. argue that – despite his explicit ambition to veer away from

authoritative ethical theories – Kitcher's project ends up advocating a privileged position in its justification of his normative stance. In other words, Kitcher commits the very mistake he tried to avoid, and this oversight will compromise the immunity to the naturalistic fallacy that he believed separated his theory from others.

O'Connor et al. point out that although pragmatic naturalism claims to expel moral authority, a privileged ethical position begins to emerge in Kitcher's normative story. The basic notion Kitcher has presented us with in his normative stance is that we should return to the conversation of our ancestors. We have avoided ethical authority, Kitcher claims, because the question is no longer about discovering ethical truth by means of religious revelation or philosophical reflection. Rather, the question has been reformulated to an if-then statement. If we want to continue the ethical project, then we must uphold the original function. In that case, it is clear that we must choose whatever path brings us back to the original function.

However, as O'Connor et al. argue, a problem arises when we attempt to resolve conflicts between derived functions. Kitcher answers that the remedy is found through his normative stance. As noted by O'Connor et al., this stance relies on two constraints: coherence and continuity. Coherence seems an obvious value – we could not succeed at any aim without coherence. But, as O'Connor et al. note, "bare coherence is presumably easily achieved by normative stances and is in fact achieved by any number of simplistic and strident ethical codes under which few if any of us would care to live" (O'Connor et al., 2012, p. 7). What Kitcher's normative stance calls for is something to direct coherence so that it maintains the ethical project as it has emerged from the original function. We find this in continuity. Continuity ensures that when derived functions conflict, we choose that function which is continuous with the original function.

The criticism here is that Kitcher seems to be treating continuity as a normative virtue rather than as a practical necessity (O'Connor et al., 2012, p. 8). As a practical necessity, continuity seems unproblematic. In fact, as a descriptive element, continuity would operate well within the spirit of Kitcher's genealogical and metaethical accounts. The problem with including continuity in his normative account is that it necessarily privileges those derived functions which lead us back to the original function, without justifying why we should be upholding the original function in the first place. Kitcher can tell us how the ethical project has evolved and how to best return to it, but he cannot tell us why we should return to it. In the end, pragmatic naturalism falls to the same objection it took pride in avoiding: the naturalistic fallacy.

O'Connor et al. offer a diagnosis for what went wrong. They write that Kitcher has successfully avoided the grounding of ethics in the "benediction of a powerful being, or the structure of rationality itself, or any such Archimedean point" and rightly separated ethical insight from "processes like scientific discovery by which we learn about persisting and independent features of the external world" (O'Connor et al., 2012, p. 10). In doing so, he has taken the authority from religion and philosophy and recognized it in the participants of the ethical project – the humans, who experience it. However, Kitcher has failed to escape the need for some transcendental backing for ethical authority. In handing the power of the ethical project to humans, Kitcher ensures the normative authority of the original function. This makes the original function a proxy for a transcendental backing of ethical authority.

Kitcher has shaped the course of naturalism in an important way: his genealogical account has made important headway in our understanding of how the ethical project evolves. However, we still have not grounded a normative stance absent a transcendental ethical authority. We are told we "ought" to return to this privileged authority, the original function,

based on a descriptive account of what the ethical project "is" and how it evolved. On the ground, on a human level, Kitcher shuffles metaethical terms around so that they have new meanings, and perhaps it begins to look like we can get an "ought" from an "is" in this new picture. But when we look closer, one of these terms is not as natural as it appears: the original function is being used as a proxy for an external moral authority, dropping in from the outside, in a convenient move to save the naturalistic story. The Humean challenge, therefore, has not been met.

The challenge for naturalists, as posed by O'Connor et al., is to learn to live without a transcendental backing of ethical authority. They write,

> We want naturalism to give us a sufficiently clear-eyed view of the status, role, and functioning of the substantive values we presently hold to be willing to defend them while recognizing that no such transcendental backing nor even a proxy for one grounded in the history or origins of the ethical project itself is ultimately possible. (O'Connor et al., 2012, p. 11)

Instead of appealing to a transcendental ethical authority, O'Connor et al. challenge naturalists to learn to live without such a backing, and to deal with first-order ethical claims head-on. We should be able to explain and discuss conflicts in values without claiming that our own moral views are privileged in this authoritative way. As O'Connor et al. write, "We thus seek a naturalistic explanation of the ethical project that allows us to look its origins straight in the eye without losing any of our enthusiasm for carrying it out" (O'Connor et al., 2012, p. 11). In the next section, I introduce another scientifically-based ethical account and examine whether it can meet O'Connor et al.'s challenge.

2.2 Joshua Greene's *Dual Process Theory of Moral Judgment*

Pragmatic naturalism failed to establish a normative ethical stance, but it did a lot to reshape how we think about the evolution of ethics. The question we may ask now is, can a naturalistic account of ethics tell us how we ought to live – without privileging an ethical authority? In this section, I present another naturalistic theory which I believe mirrors Kitcher's in important ways. On the surface, these theories do not seem to share much ground. Kitcher's theory looks at the evolution of our ethics, imploring us to return to the original function of the ethical project whereas the next theory I present, Joshua Greene's dual process theory of moral judgment, appeals to our modern human psychology, imploring us to follow our rational judgments. Importantly, although these two accounts ground themselves in different areas of the ethical experience, they both do so by privileging an ethical authority by proxy.

Greene proposes a theory which he claims meets the naturalistic fallacy without any "illicit is/ought border crossings" (Greene, 2014, p. 696). Greene attempts to show that he can use descriptive, scientific information to justify a normative claim. Although this novel approach gains some ground in the way Greene intends to use it, I argue that it sits atop a shaky foundation. That is, before Greene begins his case for evading Hume's challenge, he has already committed a critical conflation of two types of judgments: those that enhance fitness and those that produce morally correct actions. If I am correct, then Greene privileges fitness-enhancing judgments as an external authority and thereby has no justified normative claim. In other words, much like Kitcher, Greene is treating fitness-enhancing as a normative virtue rather than as a practical necessity.

Below, I Greene's theory, describing the dual processes and their neural underpinnings. Next, I present Greene's theory of moral judgments, matching each process with the judgments they produce. At this point, Greene will have only put forward a descriptive theory of moral

cognition. So, my next step is to briefly explain how Greene thinks he is able to avoid the naturalistic fallacy. Here, we'll see that Greene's solution to avoiding Hume relies on a consequentialist framework which conflates fitness-enhancing judgments with moral judgments. Without that framework, I conclude, Greene's dual-process theory fails to meet Hume's challenge.

Greene's dual-process theory is best described by analogy. In his 2014 paper, he likens the brain to a camera with automatic and manual settings. We have highly flexible automatic settings which allow us to react quickly to familiar conditions. Just as a camera has settings for "portrait" or "sports" modes, our brains have settings that are optimized for responding to situations we are familiar with due to our evolutionary past, cultural environment or personal experience. We rely heavily on these automatic settings, but when cast into an unfamiliar environment, they are insufficient.

During these unfamiliar times, we feel the urge to take control of the camera and direct our focus deliberately. We utilize a "general-purpose reasoning system, specialized for enabling behaviors that serve long(er)-term goals," which Greene calls manual mode (Greene, 2014, pp. 696-7). Here, we can guide our behavior through reasoning. Unlike our automatic settings, which mostly happen without our awareness, our manual mode involves what we usually consider to be "thinking."

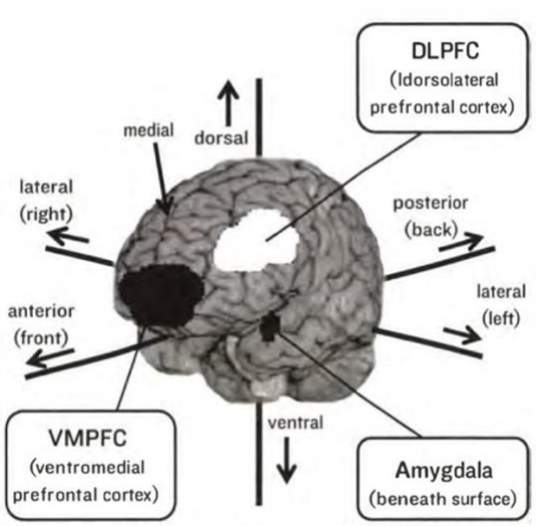
The dual-process theory is supported by recent findings in neuroscience.[8] Greene writes, "Brain regions such as the ventral striatum and the ventromedial prefrontal cortex (VMPFC)

[8] Greene's interpretation of his neuroscientific findings are not uncontested; some researchers in the field are skeptical that a dual-process view is the right interpretation. For example, vanBavel, FeldmanHall and Mende-Siedlecki (2015) argue against a dual-process theory of moral cognition, showing instead that there is a widely-distributed network of regions in the brain which underly moral judgment. They propose a switch from dual-process models of moral

produce the automatic response favoring now and enable this response to influence behavior" (Greene, 2014, pp. 697-8). The amygdala, known for its key role in emotional response, has direct connections to the VMPFC and is critical for automatic responses. Conversely, brain regions such as the dorsolateral prefrontal cortex (DLPFC) enable a more controlled response that, depending on the situation, can favor later. These responses are vital for coordinating manual mode thinking (Greene, 2014, p. 698).

Figure 2.2

Greene's Neuroanatomy of Moral Judgment



Note: A 3-D brain image highlighting three of the brain regions implicated in moral judgment. (Greene, 2013, p. 123) Used with permission of the Penguin Group.

There are a few points of disanalogy between camera operations and moral judgment, which Greene is careful to address. First, unlike a camera in which one setting precludes the other, the brain's automatic settings are always "on," even when manual mode has taken the

cognition to dynamic system models. For the purposes of this chapter, this debate does not need to be settled. Even if Greene's data were undisputed, it still would not be enough to help his theory avoid the naturalistic fallacy.

controls. Second, the settings of a camera can function independently, but the brain is different: automatic settings can happen without a manual mode, but an animal cannot have a manual mode without also having automatic settings. Lastly, the brain's automatic settings are not innate or hardwired as in a pre-programed camera. Automatic settings can be acquired or modified through cultural learning and individual experience (Greene, 2014, p. 698). The modifiability of the automatic settings should not be overstated, though – modification of an automatic setting would require processes, e.g. conditioning, that occur at a lower-level than the cognitive deliberation available to the manual mode.

So far, we have seen that the dual-process theory can explain the brain's ability to respond both in an automatic, emotional manner, as well as a controlled, calculated manner. Like a camera, the automatic settings are efficient but also inflexible, while the manual mode is flexible but time-consuming. As Greene writes, this notion is considered the "central tension in cognitive design between efficiency and flexibility" (Greene, 2014, p. 699). As it happens, this central tension also emerges in moral judgment. As Greene sees it, our brains are constantly making trade-offs in terms of efficiency and flexibility, determining which course of action is optimal for our particular circumstances. Because we respond to our environment in such a way, it seems likely that our moral judgments would follow suit; we should see them as made using similar trade-offs.

To investigate how moral judgments line up with the dual processes, Greene designs an experiment that would examine what areas of the brain are active when different moral decisions are made. Previous work by Antonio Damasio (1994) had established neural correlates of emotion and connected them to decision making. Greene hypothesizes that similarly, an fMRI would show the neural correlates of moral judgments. The inspiration for Greene's study came

from a well-known thought experiment, the Trolley Problem, which is commonly understood to demonstrate the difference between two different types of moral theories: deontology and consequentialism. Because these two theories have long been considered diametrically opposed to one another, they provide a good starting-place for a dual-process investigation.[9]

The Trolley Problem, developed by Philippa Foot (1967), aims to expose the difference between deontological and consequentialist intuitions.[10] As the story goes, a trolley is on course to kill five innocent people. If you act, you can send it off course but at the expense of killing one innocent person. A deontologist, following Kant's Categorical Imperative, is said to be committed to not killing the innocent person, as we would not want all rational people to kill an innocent person.[11] A consequentialist is said to be committed to cost-benefit reasoning and would therefore make the decision to save five lives by sacrificing one, for a net benefit of four lives.

Bringing the Trolley Problem into the laboratory, Greene and colleagues, as well as several other researchers, have gathered much data to support the dual process theory of moral judgment. They have conducted cognitive and neurocognitive studies to explore what is

---

[9] Whether or not the following understandings of deontology and consequentialism adhere to Kant's, Bentham's, or others' foundational accounts is not a matter I discuss in this paper. The versions of deontology and consequentialism I am working with are those interpretations used by Greene. Insofar as they refer to popular interpretations, and this is a theory involving moral psychology, they are important to consider here.

[10] In this chapter, I use "consequentialist" rather than "utilitarian" because it is broader and more inclusive. Although my use matches the use by Greene (2014), some of the empirical studies I cite use the term "utilitarianism." I stay consistent with using "consequentialism," so there is a mismatch of terms but the context and interpretation of empirical findings should be consistent.

[11] Greene notes that his usage of "deontological" and "consequentialist" judgments doesn't always align with philosophical usage, and qualifies them as "characteristically" deontological or consequentialist. It is important to note that the focus here is on *behaviors* and their justifications rather than commitment to an ethical school. It is irrelevant whether one who makes a characteristically consequentialist judgment commits herself in any way to consequentialism as a theory. For further explanation, please see Greene (2014), p. 699.

happening in the brain when people make consequentialist and deontological judgments in Trolley-like problems. For now, I note that there are different versions of the Trolley problem, and they vary in the responses they elicit; that is, some variations of the Trolley problem tend to induce consequentialist responses while others tend to induce deontological responses. This variation, on which I elaborate in Chapter 3, allows researchers to elicit consequentialist or deontological responses, and then look at the brain activity that corresponds. Ultimately, these researchers find that consequentialist judgments are associated with less activity in emotion-linked areas of the brain than are deontological judgments.

A plethora of studies contribute to this finding; I highlight a small but significant handful here. Studies show that in moral dilemmas where participants tend to make deontological judgments, areas of the brain that are associated with emotion, e.g., the medial prefrontal cortex, parts of the VMPFC and the amygdala, are more active (Greene et al., 2001; Greene et al., 2004).[12] This tells us that generally speaking, people make more deontological judgments when they're experiencing emotion.

There are, however, a few special cases in which certain participants make more consequentialist judgments in the same types of dilemmas that typically elicit deontological responses. Participants with damage to the emotion-linked areas of the brain, e.g., patients with frontotemporal dementia (Mendez et al., 2005) or ventromedial prefrontal lesions (Koenigs et al., 2007; Ciaramelli et al., 2007), are more likely to make consequentialist judgments than control groups.

[12] Greene et al. (2001) became the subject of controversy after McGuire, Langdon, Coltheart and Mackenzie (2009) called into question the interpretation of their fMRI findings. Greene (2009) addresses McGuire et al.'s (2009) refutation, carefully agreeing with part of their complaint while maintaining the integrity of the dual-process theory of moral judgment. Here, I have included parts of Greene et al. (2001) that do not turn on the now-dismissed interpretation.

Another group of participants show similar outcomes to those with damage to emotion-linked brain regions: participants who have more positive emotion. Valdesolo and DeSteno (2006) use an experimental manipulation to induce positive emotion in participants. Compared to a control group, participants with environment-induced feelings of positivity are more likely to make consequentialist judgments than control groups (Valdesolo & DeSteno, 2006). From this finding and the findings about patients with damage to emotion-linked areas of the brain, the message is: less negative emotion is associated with more consequentialist judgments.

Studies have also shown that consequentialist judgments are influenced by controlled cognitive processes. Using fMRI data, Greene et al. (2004) find that when participants make consequentialist judgments in moral dilemmas that typically arouse emotion-linked areas of the brain, they show more activity in areas of the brain that are associated with cognitive control and abstract reasoning, e.g., the DLPFC and the anterior cingulate cortex. In other words, participants seem to be using cognitive control and abstract reasoning to override their emotional responses, allowing them to make consequentialist judgments.

Because this fMRI data is solely correlational, however, Greene et al. (2008) set out to establish a causal relationship between controlled cognitive processes and consequentialist moral judgment (Greene et al., 2008, p. 1147). In this 2008 study, the researchers presented participants with difficult moral dilemmas which, like the 2004 study, typically arouse emotional responses. They increased the cognitive load of the (non-control) participants by giving them tasks, e.g., a digit-search task where participants hit a button when they detect a certain number, to complete simultaneously. Greene et al. (2008) found that cognitive load selectively increased Reaction Time (RT) for consequentialist judgment. In other words, for participants who made consequentialist judgments, those that performed the cognitive load task took longer than the

control group to give their judgment. This effect was not present for those participants who made deontological judgments, and in fact there was a non-significant decrease in RT when compared to the control group. Greene et al. (2008) conclude that this is direct evidence that controlled cognitive processes are at work in consequentialist moral judgments.

Greene (2014) uses these findings to support what he calls the c*entral tension principle*. This principle states that "characteristically deontological judgments are preferentially supported by automatic emotional responses, while characteristically consequentialist judgments are preferentially supported by conscious reasoning and allied processes of cognitive control" (Greene, 2014, p. 699). That is, Greene proposes that we generally make different moral judgments according to neural activity which is stimulated by environmental circumstances and our interpretation of those circumstances. Greene carefully admits that our neural activity is complex, and we surely have much to learn about neuroscience and behavior, so it would not be surprising if a particular region of the brain turns out not to work in the way he and others have presumed. However, Greene stands by the general idea that our moral theories exhibit the difference in outcomes between these dual processes.

Different Story, Same Problem

Recall that Greene's overall goal in his (2014) paper is to show that descriptive, scientific evidence can be used to justify a normative claim without committing any "illicit is/ought border crossings." So far, we've learned from Greene that we may be able to give a neuroscientific explanation of how we make moral judgments. But this is still a far cry from justifying a moral theory based on scientific evidence.

Greene thinks he can avoid Hume's challenge by taking a "direct route" for the moral implications of the dual-process theory. At first, his solution seems plausible. It would require us to buy into a variant of Hume's challenge, but it does seem to gain some ground in the way Greene intends. However, as soon as we see that this structure is the foundation upon which he makes his normative claim for consequentialism, it will be clear that a deep mistake has been made. In the foundation of his theory, Greene has conflated two very different kinds of judgments: those that enhance fitness and those that produce morally correct actions. I argue that because of this conflation, Greene is unable to show that scientific evidence can lead us toward consequentialism. In the end, Greene treats fitness-enhancing judgments as a normative virtue. Thus, much like Kitcher does for the original function, Greene privileges fitness-enhancing judgments as a proxy for an external moral authority.

Importantly, Greene never claims that scientific evidence – specifically moral psychology – alone can bring us to a normative conclusion. Instead, he proposes that it can be "normatively significant." He writes, "Moral psychology matters, not because it can generate interesting normative conclusions all by itself, but because it can play an essential role in generating interesting normative conclusions" (Greene, 2014, p. 711). He claims that he does not violate the Humean challenge because he does not base normative ethical claims purely on descriptive evidence. Below, I give an example of this move, followed by an examination of whether Greene is indeed addressing the Humean challenge.

The direct route for showing that scientific evidence can play a normatively significant role in generating interesting normative conclusions begins with asking an open normative question. It is important that the question is open because, by answering it, moral progress can be claimed. To answer the open question, both scientific evidence and an uninteresting normative

claim work together to form a new conclusion which answers the question. That is, we're making empirical discoveries about human behavior; this scientific information can be combined with an already-existing and relatively uncontroversial normative assumption (the "uninteresting normative claim"), and that combination will give us a substantive normative conclusion.

Let's pause for an example from Greene. Consider the open normative question: *Ought we condemn all incestuous behavior?*[13] Any answer here will likely be, at least immediately, controversial. This is exactly the type of question Greene believes empirical evidence from moral psychology can help resolve. He writes, "the inclination to condemn incest of all kinds is based on an emotional response whose function is to avoid producing offspring with genetic diseases" (Greene, 2014, p. 712). Those who are strong in their judgment against all incest are likely using their automatic settings, which serve a fitness-enhancing function of minimizing genetic disease. We stated this question as "open" though, which is to say that not everyone has the same response. Some might object to the categorical nature of the statement, suggesting that surely there must be some scenarios in which condemnation is not the appropriate response.

If we want to make some progress toward a new normative answer, we'll want to combine the empirical evidence about this being an emotional response with an uninteresting normative claim: If genetic diseases are not a concern, we should not rely on our emotional responses to answer this question. We can imagine a scenario where siblings separated at birth meet and develop an intimate relationship. Let's say that even though they find out they're related, they take every precaution and use a permanent form of birth control to prevent pregnancy. If offspring are not possible, then our emotional responses are firing for no reason.

[13] It actually doesn't matter that the reader agrees *this* question is open. One must only allow that there is *some* open normative question in order to follow Greene's proposal.

Together, the scientific evidence and this uninteresting normative assumption give us a new, substantive normative conclusion: We ought not to condemn all incestuous behavior. We just answered a difficult ought question by appealing to [1] scientific ("is") information and [2] an easier "ought" assumption than we started with. For Greene, this amounts to moral progress by way of empirical evidence. I will examine this move further, but first let's consider whether he is meeting Hume head-on.

Greene promises not to commit any is/ought sleight of hand in his proposal that neuroscience and cognitive science can have implications for ethics. So let's ask, is it a fair move as far as Hume is concerned, to use both an "is" and an "ought" to infer a stronger "ought"? Adding to the quotation in the introduction of this paper, Hume is perplexed by moral theorists who spend quite a bit of time in the "ordinary way of reasoning" either establishing the existence of a God or describing observations of human behavior. Then, out of nowhere, every proposition the theorist writes seems to be an "ought" or an "ought not." Hume expresses concern: we should be weary of a moral theory in which a new type of proposition appears as a "deduction from others, which are entirely different from it" (Hume, 1739/2000, T3.1.1.27).[14] Hume only speaks to inferences in which exclusively descriptive propositions lead to exclusively normative conclusions. He does not say anything about inferences in which there are normative and descriptive propositions leading to a normative conclusion. By this measure, Greene seems to be in the clear: he is not violating Hume's challenge (yet).

[14] After this section in Hume's *Treatise*, he concludes that morality is not discoverable by reason alone. It would be interesting to further investigate Hume's idea of "moral sense" and how that may be thought of in light of moral cognition. For the purpose of this paper, though, I am restricting the scope of the Humean challenge to what Hume states in T3.1.1.27, which also overlaps with the naturalistic fallacy and the is/ought gap.

There is something else going on here: a shifting of the ground in a way that Hume just did not see coming. Although Greene explicitly states in his paper that he is leaving aside questions about metaethics, this move is reminiscent of Kitcher's metaethical redefinitions: it is an answer not to Hume's original theory, but to an alternate version.[15] For each Kitcher and Greene to claim victory over Hume, we need to take significant steps away from the original Humean challenge. So, in each case, we might say, something interesting is being done and it is related to the Humean challenge – to what degree may be debated.

In the end, the cleverness of Kitcher's metaethical redefinitions did not matter for pragmatic naturalism; it still violated the Humean challenge by privileging the original function as an external ethical authority. Following a similar fate, we will see that the cleverness of Greene's quest for normative significance will not matter for the dual-process theory. Later, we will examine a deep mistake that Greene has made in the very structure of the direct route.
So far, Greene has shown us how our brains act when we make moral judgments. He has combined this information with uncontroversial normative claims to help us make stronger normative claims, as if to say: If we understood ourselves better, we'd see why we're reacting in this way, or even: If we just slow down and think more, we can make judgments that improve our lives and the lives of others. It is critical, though, to point out that Greene has not actually shown that the direct route leads to any normative significance. That is to say, he has not shown us that his theory leads us to better moral conduct.

[15] Recall that Kitcher's defense against the naturalistic fallacy depended on a switch in terminology of "truth" and "progress." This switch was a radical departure from conventional usage, and it is questionable whether Kitcher is even playing the same game as Hume, let alone evading Hume's challenge. Likewise, Greene is radically departing from the traditional view of Hume's challenge by looking for normative significance rather than purely an "ought" from an "is."

This is precisely where Greene fails Hume's challenge: he draws conclusions of normative significance from judgments which are nothing more than fitness-enhancing. Below, I describe Greene's two-part push for consequentialism and then I argue that its very foundation rests on a conflation of judgments that produce moral conduct and judgments that enhance fitness.

Greene's first step toward his consequentialist conclusion comes through the direct route: having scientific information about our moral psychology can significantly impact how we make normative claims. Our emotional responses to moral dilemmas can be reviewed and our automatic settings kept in check. When we take the time to use our manual mode, our judgments become more nuanced. Though one may be tempted to assume the manual mode is always better than the automatic settings, Greene believes that "automatic settings and manual mode are respectively better and worse at different things" (Greene, 2014, p. 714). Just like for other types of judgment – and for cameras – there is a trade-off between efficiency and flexibility. So far, we are left with a sort of pluralism: some situations will call for automatic settings, and some for the manual mode. The question Greene takes on is: Which situations call for which settings?

In answering this question, Greene sets up the second step of his push toward consequentialism: he makes a case for how to prescribe each type of response. This requires caution, as Greene is reluctant to make any overtly metaethical claims. His answer is reminiscent of Kitcher's normative guidance in that it is a practical solution based on behaviors and outcomes rather than an allegiance to some external authority. As I show, however, it also follows Kitcher's lead in that it ultimately uses a proxy for an external authority.

Moral judgments often rely on automatic settings when a situation is familiar to us, meaning that it has been shaped by trial-and-error experience. There are three mechanisms

known to give us this sort of experience: genetic transmission, cultural transmission, and learning from personal experience (Greene, 2014, p.714). Some examples are: genetic transmission of our ancestors' experiences predisposes us to fearing snakes; cultural transmission of experience causing us to fear guns even if we've never been harmed by one; personal experience teaching us to fear hot stoves after we've touched them. When we are in a moral situation that evokes this repertoire of experiences, our automatic settings kick in and we act efficiently. This is not necessarily a bad thing – in fact, our evolutionary success has no doubt been shaped by our automatic responses to familiar settings. There may be times where we override these automatic settings – for example, overcoming a fear of snakes through careful conditioning. Greene's point here is not that we must exercise our manual mode in familiar conditions, but rather, that we cannot exercise our automatic settings in unfamiliar conditions.

Employing our automatic settings when we are in unfamiliar conditions is asking for a cognitive miracle. Greene writes, "For one of our automatic settings to function well, its design must be informed by someone's trial-and-error experience" (Greene, 2014, p. 714). If we have no genetic, cultural or personal experience in a situation, it is unfamiliar and our automatic settings have no basis from which to operate.[16] So, Greene offers the No Cognitive Miracles Principle, which states that when we are dealing with unfamiliar moral problems, we ought to rely more on the manual mode and less on our automatic settings; otherwise, we're banking on cognitive miracles (Greene, 2014, p. 715).

Greene suggests that our moral judgments ought to be guided by our scientific findings. Explicating his thesis of the significance of moral psychology, he writes:

[16] The asterisk for *unfamiliar* is used by Greene to indicate a specified technical meaning, in a similar fashion as was done for *deontology* and c*onsequentialism*.

> If we believe that we ought to rely on automatic settings vs. manual mode to different extents in different situations, and if cognitive science can tell us when we are relying on automatic settings vs. manual mode, then cognitive science gives us normatively significant information – information that can nudge us, if not propel us, toward new and interesting normative conclusions (Greene, 2014, p. 715).

It is important to note here that Greene considers practical moral disagreement to be under the umbrella of unfamiliarity. That is, if two people disagree about a response to a supposedly-familiar situation, then they are having contradictory automatic responses.[17] In this case, Greene says that at least one of them has automatic settings that are "misfiring" but there's no readily available way to determine who's having the misfire. He suggests here that we distrust our automatic settings and go into manual mode (Greene, 2014, p. 716). As we recall, according to the central tension principle, pulling into manual mode places us in consequentialist-territory.

In the end, Greene believes that the dual-process theory favors consequentialism. He writes:

> We should distrust our automatic settings and rely more on manual mode when attempting to resolve practical moral disagreements… I believe [this] favors consequentialist approaches to moral problem solving, ones aimed solely at promoting good consequences, rather than deontological approaches aimed at figuring out who has which rights and duties, where these are regarded as constraints on the promotion of good consequences. (Greene, 2014, p. 717)[18]

Greene notes that deontological philosophers certainly make use of the manual mode; it would be unfair to characterize Kant et al. as only thinking emotionally. However, Greene suggests that the work they do in the manual mode is not actually moral reasoning with the goal of figuring

[17] It might be assumed that if there is a moral disagreement, either the situation is being observed or experienced differently by each person, or the repertoire of ancestral/cultural/personal experience is different for each person.

[18] Greene believes that his view favors act consequentialism (and in some ways, rule consequentialism). For the purposes of this paper, the nature of Greene's consequentialism is not important; rather, it's only important *that* he makes a normative push for consequentialism.

out what is right or wrong; rather, it often amounts to moral rationalization where "their reasoning serves primarily to justify and organize their preexisting intuitive conclusions about what's right or wrong" (Greene, 2014, p. 718). Furthermore, consequentialism – as supported by the manual mode – is more transparent than deontology and its automatic responses. In acting consequentially, we are not committed to intuitions and acting in response to emotion. Rather, our actions can be explicated and justified in terms of the consequences they produce. Greene has made a clear case for consequentialism, but will it work?

I have already argued that the first step – the direct route – may have avoided the Humean challenge, but in a weak sense. Greene's second step, however, will not follow suit. In the foundation of his theory, Greene has conflated two very different kinds of judgments: those that enhance fitness and those that produce morally correct actions. Because of this conflation, I argue, Greene is unable to show that scientific evidence can lead us toward consequentialism. In the end, Greene treats fitness-enhancing judgments as a normative virtue. Thus, much like Kitcher does for the original function, Greene privileges fitness-enhancing judgments as a proxy for an external moral authority.

Furthermore, moral disagreement is problematic for the automatic settings because there is no way to determine who is right. Both parties are having different automatic responses to a dilemma, but we are unable to determine whose responses are "misfiring." So, Greene says, we need to go into manual mode to make this determination and this is what tips the scale in favor of consequentialism. Because of the central tension principle, we know that going into manual mode will result in more consequentialist judgments, and so from the start, Greene is advocating for a consequentialist resolution of moral disagreements. To be clear, to say that for any moral disagreement to be resolved via consequentialism is to deny progress to any other moral theory.

If, for example, a moral disagreement arose between two deontologists, then according to this theory, they should go into manual mode, where their judgments will take on a consequentialist nature. In Greene's (2014) paper, he does not mention other moral theories, e.g. virtue ethics or care ethics, but he has left us with no choice but to infer that moral disagreements in these other theories would also need to be resolved on consequentialist grounds.

Furthermore, Greene criticizes deontologists for only using the manual mode for justifying preexisting intuitive conclusions, but the case he has given for consequentialism is no better. The work done here is not actually moral reasoning with the goal of figuring out what is right or wrong either! It is limited to figuring out what will enhance fitness. What Green successfully shows is that our manual mode allows us to respond in a more nuanced, rational way to our environment. There is no moral claim here.

So, what we've seen is that, in both the pluralism-stage and the consequentialism-stage, Greene's theory makes a critical error: it conflates judgments that enhance fitness with judgments that produce morally correct actions. Automatic settings are good for familiar conditions that we're biologically, culturally or personally prepared for – and we should rely on them in those settings. Similarly, the manual mode is good for unfamiliar conditions like new situations or for moral disagreements – and we should rely on it in those settings. Greene gives the reader many reasons why this is the case: the manual mode makes us more thoughtful and less emotional; it allows us to make impartial cost-benefit reasoning and helps us to understand why personal dilemmas seem different to us than impersonal ones. But at no time does Greene argue that the manual mode produces morally correct actions more often or more accurately than the automatic settings.

Greene says we are "propelled" toward consequentialism, but this is not based on evidence that we will be making morally correct judgments; it is only based on evidence that we will be making fitness-enhancing judgments. There is still an "is/ought gap" here. Aiming to close it, Greene violates the naturalistic fallacy by privileging fitness-enhancing judgments as a proxy for an external ethical authority. Fitness-enhancing judgments would fit well within a descriptive dual process theory; they can explain our actions and why we seem to have conflicting moral intuitions. But in a normative stance for consequentialism, they operate as a proxy for a transcendental ethical authority: some outside force toward which we should aim, without any naturalistic authentication.

2.3 The Moral Learned

These two naturalistic accounts of morality show us that attempting a claim of normative significance from descriptive evidence is not a promising endeavor. I have presented only two accounts, but these accounts take very different paths. Kitcher is focused on what knowledge about our evolutionary history can teach us. He privileges the original function of our ethical project, appealing to its role in our ancestral past. Greene, on the other hand, is focused on our modern neuropsychology. He privileges our rational thought processes, appealing to their ability to enhance our fitness. However, even with different approaches, each theory ultimately violates the Humean challenge, and for the same reason: they privilege a scientific entity so that it is used as a proxy for a transcendental ethical authority.

One moral we might learn from Kitcher and Greene is that perhaps scientific theories must be pushed out of the descriptive realm in order to gain normative consequences. In reaching

to grasp onto normative consequences, the once-descriptive theory loses its footing, requiring a boost in the form of an ethical authority, either directly or by proxy. If this is the case, then scientific accounts can never support normative conclusions. As tempting as it may be for naturalists to find a scientific answer to what we ought to do, such a theory is no more than a mirage. Just when we think it is in our grasp, it disappears into thin air. Although our failed grasp may be disappointing, the reality is that this dream was never within our grasp in the first place. It's time to reconsider what science can tell us about ethics, because it seems the path of normativity is not a fruitful one.

A second moral we may learn here, and where I will invest my efforts, is that there must be other ways in which science is important for our understanding of morality. In these works, Kitcher (2011) and Greene (2014) were both attempting something like Kitcher's (2006) descriptions of Projects 2 and 4, as presented in Chapter 1. As I argued in response to Kitcher (2006), there is a better way to think of the contributions of science to ethics, viz., that science can make a practical difference in our moral deliberations. We should not attempt to squeeze normativity out of these empirical accounts; there are much better uses for our scientific knowledge. Science uncovers fascinating and important information about our moral cognition. Through Kitcher (2011), we can imagine how ethics has emerged from our evolutionary history. Through Greene, we learn how our moral judgments are supported by two different cognitive systems, each pushing toward different judgments. This information is incredibly important, not because it teaches us what we should value but because it teaches us how to uphold our values.

CHAPTER 3: An Informed Introspection

3.0 The Introspection Illusion

In the past two chapters, I have built a case for why we should use scientific information about our moral cognition in our moral deliberations. This motivation has been built on philosophical grounds: science should play a practical role rather than, say, a normative role, in ethics; also, involving science in our moral deliberations will improve our moral agency. In introducing this chapter, I add some more depth to this motivation by briefly looking to psychological research on how we think about our own cognitive processes.

Psychologist Emily Pronin has investigated two related areas that should be considered when we think about how people can make changes to their moral deliberation. The first is what she terms people's bias blind spot. People are able to recognize that biases in cognitive processes exist and that they have an impact on human judgment and inference; as it turns out, though, people are less able to recognize these biases in themselves (Pronin, 2007). In other words, we're quick to point out biases in how others think – in fact we often overestimate the impact of biases on others – but fail to acknowledge that we ourselves may be susceptible to bias. Secondly, people have a persistent and widespread tendency to heavily weigh introspection when we seek to understand ourselves; Pronin calls this tendency an introspection illusion (Pronin, 2009, p. 3). That is, when we seek to understand our own judgments and actions, we overvalue our introspective access to our conscious processes like our emotions, thoughts, and attitudes. Scientific findings from social psychology, cognitive science and neuroscience (e.g., the dual-process theory of moral judgment) have revealed that a significant portion of our judgments and

actions happens without our awareness, effort, or intent (Pronin, 2009, p. 2). A critic, then, may look at my project and say something like, Teaching people about bias will just arm them against others, and asking them to focus on their own moral deliberations will just escalate their overvaluing of introspection. This critic should be answered.

Responding to the critic's concerns will illuminate the picture I wish to create. Pronin's work points out two very real risks of my project. The first, regarding the bias blind spot, is that if people learn more about all the biases and errors involved in our cognitive processes, they'll use this information to devalue the judgments of others and remain convinced that their own cognitive processes are immune to these errors. I believe that Pronin's work is exactly the type of scientific research we should include in our moral deliberations. If I know that I may have a bias blind spot, then I should look for those blind spots. Fortunately, the science is encouraging here: Pronin and Kugler (2007) report that study participants were freed from the bias blind spot when they were educated about the fallibility of introspective evidence. So, by learning about the role of nonconscious processes in our judgments and actions, people can be liberated from their bias blind spots. This is precisely the sort of outcome that will promote our own agency – we will see how cognitive biases affect others' and our own judgments.

The second worry, the introspection illusion, is directly addressed by my project. By welcoming scientific information about our moral cognitive processes into our moral deliberations, we are moving toward what I call an informed introspection. Here, we are aware of our bias blind spots, as well as an array of other cognitive processes that reveal surprising information about our moral processes. We can incorporate this knowledge into our introspection, allowing us to change our moral deliberations in ways that are more attuned to scientific research and our own moral principles. Furthermore, and as Pronin suggests, we should

be called to better weigh our own introspective process with others' introspections, i.e., their reflections on their internal thoughts, feelings, and motives (Pronin, 2009, p. 54). In doing so, we can better balance our moral deliberations, taking into consideration our own introspective limits and how others value their own introspection.

As I continue this chapter, I ask the reader to consider different ways in which scientific knowledge can help inform our moral deliberations. Inspired by Pronin's work, I believe that we will find that learning about our moral processes can be effective (i.e., we will have a fighting chance to resist our bias blind spots) and that we will come away with an enhanced deliberative process (i.e., an informed introspection). The scientific information I discuss in this chapter involves neuroscience, cognitive science, and social psychology. The findings from these disciplines ask us to rethink how we view moral regret, as well as how we should use our factual beliefs in our moral reasoning. In Chapter 4, I continue my project by using science to suggest that we should be skeptical of our moral intuitions, and that doing so will increase coherence in our moral deliberative processes.

The three uses of scientific information in this dissertation will have notes of similarity and dissimilarity. They should all contribute to my project, meaning that each is an example of how we can use scientific information to make concrete changes to our moral deliberation. Nonetheless, these cases make different sorts of contributions. I intend for this project to be pluralistic – that is, I am open to many ways in which science can make this contribution. There are important questions and challenges that arise during the presentation of each individual case. My answers to these challenges may be generic or case-specific. However, I anticipate that even when my answers seem case-specific, they may be useful in response to challenges that arise in future examples.

3.1 Relief from Moral Regret

Greene's dual-process theory (described in Chapter 2) shows us that moral judgments which have inspired various – seemingly contrasting – moral theories are naturally within us. What is important about Greene's theory, I argue, is different from what he derives from it. After introducing us to two separate evolved systems for making moral judgments, Greene champions one over the other. As I argue in the previous chapter, he thereby commits the naturalistic fallacy. Nonetheless, at this point, I separate Greene's research from his theory. While I find his push for consequentialism unacceptable, I firmly believe there is value in the scientific facts he has brought to light.

Whereas traditional approaches to ethics have diverted philosophers into comparing one normative theory to another, Greene's empirical investigations reveal the very human nature behind both consequentialism and deontology. As described in the previous chapter, the dual process theory of moral judgments tells us that two different evolved systems are at work when we make moral judgments: consequentialist judgments involve the manual system and deontological judgments involve our automatic settings. This theory, then, can account for what may otherwise seem to be theoretically inconsistent moral behavior, where a person may behave according to one theory in one moment and another theory in the next. Rather than viewing the choice made in a moral dilemma as one that represents a person's character or the moral theory to which they ascribe, I argue that we should consider the cognitive processes at play during a person's decision.

These considerations become more evident when we examine the dual-process theory more closely. In the previous chapter, I mentioned that there are various types of Trolley-like dilemmas – some which tend to evoke more deontological (and emotional) responses than others. What is interesting to Greene and other dual-process researchers is the intrapersonal variation in moral dilemma judgments. Even though, on paper, two different dilemmas may look similar – i.e., they involve the same quantities of sacrifice and gain – people tend to respond very differently to them. In this section, I describe two versions of the Trolley Problem and two complementary scientific theories that aim to explain why we judge them so differently. Then, I offer a real-life moral dilemma in which a soldier made a deontological choice and has since suffered greatly from moral regret. I argue that the dual-process theory of moral judgment gives us good reason to reinterpret the soldier's moral regret, with the hope that this will bring him, and those of us who find ourselves in similar situations, some relief.

Trolley Dilemmas

The Trolley Problem can be structured in a way that exposes an inconsistency in moral judgments. In the traditional Trolley Problem, the switch problem, participants are told to imagine a runaway trolley is headed down a set of tracks where there are five innocent people who will die upon impact. Participants are then asked whether they will flip a switch to change the trolley's path – directing it to other tracks, where one innocent person will be killed upon impact. In response, people tend to make a consequentialist judgment: saving five lives at the cost of one is the right thing to do (Greene, 2001, p. 2105).

However, in a different version of the Trolley Problem, the footbridge problem, participants are told to imagine they are on a footbridge overlooking tracks as a runaway trolley

is headed toward five innocent people. There is a large man on the footbridge, and the participant is asked whether they will physically push him off the footbridge, onto the track, killing him but saving five lives.[19] In this case, participants tend to make a deontological judgment: killing an innocent person is wrong regardless of the consequences.

The puzzle, then, is why we make different moral judgments when the math is the same: we net four lives. Connecting this phenomenon to the discussion in the previous chapter, Greene's fMRI data indicates that the footbridge dilemma correlates with activity in areas of the brain known to be connected to our emotional responses (Greene, 2001). Conversely, the switch dilemma correlates with areas of the brain known to be connected to cognitive control and abstract reasoning (Greene, 2004). Thus, Greene theorizes that when we are thinking of the footbridge scenario, we have less cognitive control of our emotional responses, leaving us unable to go into manual mode and inhibit our automatic responses that are telling us "Don't kill people!". In other words, the footbridge leaves us in our automatic settings. When we are thinking of the switch scenario, we aren't having as much activity in our emotion-linked brain areas and so we are able to maintain cognitive control and reason through the dilemma using our manual mode.

Again, the math is the same in both the footbridge and switch Trolley Problems. However, to most people, they are fundamentally different. Greene and other Trolleyology researchers have attempted to pinpoint exactly what this difference is. Although there is no definite consensus, several studies have put forward explanations for this distinction.[20] Here, I introduce two complementary theories which have become central to explaining this

[19] The footbridge Trolley Problem was introduced by Judith Jarvis Thomson (1985).
[20] See Greene (2009) for a list of studies proposing explanations.

phenomenon. The first, introduced by Greene et al. (2001), views the distinction between the footbridge and switch as a matter of personal and impersonal dilemmas. The second, put forward by Valdesolo and DeSteno (2006), shows that responses to the footbridge scenario can be influenced by positive emotion while responses to the switch scenario are not influenced.

Greene et al. (2001) was groundbreaking. Greene and his colleagues applied cognitive neuroscience to moral philosophy, in search of neural correlates to our moral judgments. To elicit deontological and consequentialist judgments, the researchers used the footbridge and switch dilemmas to probe participants while conducting an fMRI to view their brain activity in real-time. Because fMRI data is noisy, the researchers needed more data points than the footbridge/switch distinction provides. So, they came up with what they believed to be the salient difference between the two cases: the footbridge case requires a personal action (i.e., physically pushing a person over a footbridge) while the switch case requires only an impersonal action (i.e., flipping a switch from a distance).

The researchers developed 60 practical dilemmas that reflected a personal/impersonal distinction, dividing the dilemmas into three conditions: moral-personal (e.g., footbridge), moral-impersonal (e.g., switch) and non-moral (e.g., which of two coupons to use at a store). Remarkably, Greene et al. found that in terms of the psychological processes associated with their production, judgments concerning 'impersonal' moral dilemmas more closely resemble judgments concerning non-moral dilemmas than they do judgments concerning 'personal' moral dilemmas" (Greene et al. 2001, p. 2107). In other words, deciding to flip the switch is more similar to deciding which coupon to use than it is to deciding to push a person off a bridge! In an "excellent piece of scientific detective-work" the personal/impersonal distinction was reanalyzed by McGuire et al. (2009), undermining the interpretation by Greene et al. (2001)

(Greene, 2009). Their results were found to be an artifact of several of the dilemmas they had developed to match the footbridge/switch cases. Nonetheless, subsequent studies have continued to build on the personal/impersonal distinction as a defining difference between dilemmas that induce deontological versus consequentialist responses. Greene et al. (2009) examines the use of personal force during moral dilemmas. They find that harmful actions are judged as less morally acceptable when an agent applies personal force (Greene et al., 2009, p. 369). Furthermore, Moore et al. (2008) redesigned the Greene et al. (2001) experiment, finding support for the personal/impersonal distinction, and Moore et al. (2011) found cross-cultural evidence for this distinction, finding it in populations from the U.S. and from China. So, although the road has not been smooth, it does seem that the personal/impersonal distinction is at least part of the scientific explanation for why we think of the switch and footbridge dilemmas as being fundamentally different.[21]

Another idea about the difference between these two Trolley Problem variations is that one of them, viz. the footbridge scenario, is susceptible to environmentally-induced changes in affect. In other words, how we process the footbridge dilemma can change, depending on whether something in our environment changes our mood. Valdesolo and DeSteno (2006) showed 79 participants either a positive 5-minute comedy video clip (from *Saturday Night Live*) or a 5-minute neutral video clip (from a documentary on a Spanish village). Then, the participants were given both the footbridge and switch dilemmas to judge. The researchers found that the participants who watched the comedy clip both reported a more positive affective state

[21] The inclusion of this back-and-forth is not meant to detract from the strength of the overarching dual-process theory. In fact, it is an example of the scientific rigor that has been applied to the theory and as such, is a testament to its strength.

and were 3.8 times more likely to make a consequentialist judgment in the footbridge dilemma (Valdesolo & DeSteno 2006, p. 477). There was no effect for responses to the switch dilemma.

Valdesolo and DeSteno (2006) report that these findings "demonstrate that the causal efficacy of emotion in guiding moral judgment does not reside solely in responses evoked by the considered dilemma, but also resides in the affective characteristics of the environment" and that it is clear that "a skilled manipulation of individuals' affective states can shape their moral judgments" (Valdesolo & DeSteno, 2006, p. 477). Remarkably, Valdesolo and DeSteno have shown that when people are judging a footbridge dilemma, their emotional state is crucial to their judgment. They are not deciding the dilemma solely on its descriptive merits. When a person feels positive emotion, they make a more consequentialist judgment. Combined with the evidence presented in the previous chapter about the effect of negative emotion on judgments in footbridge dilemmas, the message is: when we're considering a moral dilemma like the footbridge, a negative emotional state will likely elicit a deontological judgment and a positive emotional state will likely elicit a consequentialist judgment.[22] What's more, Valdesolo and DeSteno have shown that a relevant affective state can be manipulated by a person's environment. That is, the researchers were able to alter a person's emotional state effectively enough to garner a different moral judgment.

Notably, this manipulation does not work in the switch dilemma. So, Valdesolo and DeSteno seem to have uncovered an important aspect of the difference between how we make moral judgments according to different brain processes. Because our affective states determine the information signals about our environment, Valdesolo and DeSteno write that the dual-

[22] In Chapter 2, I did not refer to this type of dilemma as a "footbridge" dilemma but rather as a type of moral dilemma that "typically elicits deontological responses."

process model of moral judgments suggests that our choices may be influenced by contextual sensitivity of affect (Valdesolo & DeSteno, 2006, p. 476). In other words, the dual-process theory holds that information from our environment is crucial in determining which process (i.e., manual mode or automatic settings) is active. These researchers have shown that the type of dilemma with which we are confronted is not the only piece of environmental information that determines which process is active. If the dilemma is footbridge-like, then our affective state is also important in determining our moral judgment – and crucially, this can be manipulated by our environment.

Together, the personal/impersonal distinction and the affect effect begin to show how the switch and footbridge moral dilemmas differ. There are probably additional dimensions in which they differ, and as the scientific investigation continues this picture will likely become clearer. The takeaway for my project, though, is that what seemed like a puzzling inconsistency in moral judgments before scientific investigation has become more understandable after examining the cognitive processes involved. The work of Greene and other researchers to illuminate our dual-processing of moral judgment has added depth and predictability to a moral dilemma that previously perplexed philosophers. I argue that this has significance for our moral reflective processes. In the next sub-subsection, I use a real-life example of a soldier who made a deontological choice in a footbridge-like dilemma and has lived with deep moral regret ever since.[23] After examining the dilemma this soldier faced and with a new understanding of the cognitive processes likely involved, I believe this soldier is warranted a new appraisal of his moral guilt.

[23] This real-life footbridge dilemma was discussed in Sandel (2009). Sandel uses the example normatively, to discuss how to reason our way through disagreements about justice whereas I am using it descriptively, to exemplify our dual-process system of moral judgment.

An Impossible Choice

In Afghanistan, on the morning of June 27, 2005, U.S. Navy SEALS Lieutenant Mike Murphy, Petty Officer Matthew Axelson, Petty Officer Danny Dietz and Petty Officer Marcus Luttrell were given orders to capture or kill Ahmed Shah, a leader of a Taliban force who was considered responsible for several lethal bomb attacks on U.S. forces. Though they were experienced soldiers, Luttrell recounts in his book, Lone Survivor, that they struggled with the rules of war. Luttrell writes,

> We have an extra element of fear and danger when we go into combat against the Taliban or al Qaeda – the fear of our own, the fear of what our own navy judge advocate general might rule against us, the fear of the American media and their unfortunate effect on American politicians (Luttrell, 2007, p. 171).

While Luttrell and other American soldiers must abide by the Geneva Convention, the Taliban and al Qaeda did not engage by the same rules. This was stressful for the soldiers, knowing that their in-the-moment actions may be deemed criminal by a court after the fact. Luttrell describes encounters with Afghan men in remote mountains where it is unclear whether the men are armed combatants or unarmed civilians until they attack the soldiers with rockets.

Fears of coming under fire from the Taliban were compounded by fears of later facing American courts haunted Luttrell and his team as they prepared to embark on their mission. As they gathered their explosives and equipment, and waited for their helicopter transport, the soldiers began to feel unsettled. The maps they had seen of their destination showed few places to hide. Luttrell writes,

> Every one of the four of us, Mikey, Axe, Danny, and me, made it clear, each in his own way, that we did not have a good feeling about this. And I cannot describe how unusual that was. We go into oops areas full of gung ho bravado, the way we're trained – Bring 'em on, we're ready! No SEAL would ever admit to being scared of anything. Even if we were, we would never say it. We open the door and go outside to face the enemy,

> whoever the hell he might be. Whatever we all felt that night, it was not fear of the enemy, although I recognize it might have been fear of the unknown, because we really were unsure about what we would encounter in the way of terrain (Luttrell, 2007, p. 188).

The soldiers went on their way navigating through the mountains in the dark of night. After a grueling seven-hour hike, the soldiers began to look for a place to settle. There was no good option. Luttrell writes that it was "every frogman's dread, an operation where the terrain was essentially unknown and turned out to be as bad as or worse than anyone had ever dreamed" (Luttrell, 2007, p. 197). The men carried on, and carried with them a deep feeling of unease.

As the soldiers assessed their surroundings, an Afghan man carrying an ax startled Luttrell, who swiftly pointed a gun at him. The man seemed surprised to see the soldiers and followed their gestured instructions and dropped the ax. At this point, the four soldiers were assessing whether the Afghan man is a Taliban combatant or a civilian. Then came along another man, a teenage boy and about a hundred goats. The Afghans appeared to be goatherds and in their limited English, denied any affiliation with the Taliban. Luttrell gave the teen a power bar snack as all four soldiers debated what to do about these unarmed civilians. Luttrell writes,

> The question was, What did we do now? They were very obviously goatherds, farmers from the high country. Or, as it states in the pages of the Geneva Convention, unarmed civilians. The strictly correct military decision would still be to kill them without further discussion, because we could not know their intention. How could we know if they were affiliated with a Taliban militia group or sworn by some tribal blood pact to inform the Taliban leaders of anything suspicious-looking they found in the mountains? And, oh boy, were we suspicious-looking (Luttrell, 2007, p. 202).

The soldiers did not have any equipment to temporarily disable the goatherds – they had dropped their heavy load before their trek up the mountain. The only two options were to either kill the goatherds or let them go. Axe said they should kill the goatherds because it was too risky to let them go. Danny said he'd do whatever everyone else voted to do. Mikey reasoned through their release:

> If we kill them, someone will find their bodies real quick. For a start, these fucking goats are just going to hang around. And when these guys don't get home for their dinner, their friends and relatives are going to head straight out to look for them, especially for this fourteen-year-old. The main problem is the goats. Because they can't be hidden, and that's where people will look. When they find the bodies, the Taliban leaders will sing to the Afghan media. The media in the U.S.A. will latch on to it and write stuff about the brutish U.S. Armed Forces. Very shortly after that, we'll be charged with murder. The murder of innocent unarmed Afghan farmers (Luttrell, 2007, p. 203).

Luttrell was paralyzed by Mikey's reasoning and said they needed advice. They attempted to call their headquarters, but the comms system was down. At that moment, the four soldiers realized they were on their own. If they killed the Afghans, they could not do so in plain sight; they would need to cover it up. Otherwise, they'd have the Taliban looking for them in no time. So the decision was either to kill unarmed civilians and hide their bodies, or to let them go. Luttrell writes about his state of mind in the moment,

> If this came to a vote, as it might, Axe was going to recommend the execution of three Afghans. And in my soul, I knew he was right. We could not possibly turn them loose. But my trouble is, I have another soul. My Christian soul. And it was crowding in on me. Something kept whispering in the back of my mind, it would be wrong to execute these unarmed men in cold blood. And the idea of doing that and then covering our tracks and slinking away like criminals, denying everything, would make it more wrong. To be honest, I'd have been happier to stand 'em up and shoot them right out in front. And then leave them. They'd just be three guys who'd found themselves in the wrong place at the wrong time. Casualties of war. And we'd just have to defend ourselves when our own media and politicians back in the U.S.A. tried to hang us on a murder charge (Luttrell, 2007, p. 205).

All four soldiers disliked the sneaky option, and Luttrell attributes this to their Christian values. They understood the correct military decision was to openly kill the goatherds, accepting any judicial consequences, because the risk of letting them go was too high for the mission. In the end, though, neither killing-option felt right to Luttrell. He cast the final vote in favor of setting the goatherds free.

Soon enough, the soldiers' worst fears materialized as they realized their location had been compromised. The man they were tracking, Shah, and his Taliban fighters opened fire. Axe, Danny and Mikey were killed in battle, as were 16 more soldiers when their MH-47 rescue helicopter was shot down. Luttrell is the sole survivor from the attack.

As the only decision-maker that lived through that tragic day, Luttrell's regret has been a constant presence in his life ever since. Luttrell reflects:

> It was the stupidest, most southern-fried, lamebrained decision I ever made in my life. I must have been out of my mind. I had actually cast a vote which I knew could sign our death warrant. I'd turned into a fucking liberal, a half-assed, no-logic nitwit, all heart, no brain, and the judgment of a jackrabbit. At least, that's how I look back on those moments now. Probably not then, but for nearly every waking hour of my life since. No night passes when I don't wake in a cold sweat thinking of those moments on that mountain. I'll never get over it. I cannot get over it. The deciding vote was mine, and it will haunt me till they rest me in an East Texas grave (Luttrell, 2007, p. 206).

There seems to be no hiding from this regret for Luttrell. What I offer here, however, calls for a reconsideration of the moral regret he faces. In the next subsection, I show that when we consider the dual-process theory, our view of Trolley Problem-like dilemmas will change. Importantly, I argue that this view of the cognitive events can offer us some relief from moral regret in a situation like Luttrell's.

Reinterpreting Moral Regret

Let us summarize what happened to Luttrell: he was in a footbridge-like moral dilemma, tasked with choosing between murdering innocent civilians in cold blood or letting them go free, risking his own personal safety and the safety of his fellow soldiers. He chose the latter, and the worst-case scenario happened. Now, let us summarize Greene's dual-process theory of moral judgment: In a footbridge-like moral dilemma, our automatic processes are favored. People tend

to make deontological choices when they are experiences negative emotions. In this light, we have a viable explanation for Luttrell's choice. From his memoir, we know that he was experiencing a high level of negative emotions – he was carrying the weight of a bungled operation and felt the gravity of his Christian values. We may pause and say, "Of course Luttrell made the deontological choice – his automatic system was in charge, and we make deontological choices when our automatic system is in charge!" Although I am sympathetic to this response, I want to explore a more nuanced approach – one that does not turn us into hard determinists about our moral actions.

Before considering the nuanced approach, let's examine the consequences of absolving Luttrell's moral regret by pointing to Greene's work as an exculpation of Luttrell's moral behavior. Thinking of the dual-process theory as offering a cognitive explanation of our moral behaviors invokes a cognitive monster. This term, borrowed from John Bargh (1999), refers to the worry that a significant amount of our behavior happens outside of our conscious processing. If we are not able to direct this behavior, it seems questionable to say that we could be morally responsible for it. To contextualize this worry in Luttrell's case, consider the following argument: Luttrell was not in charge of which process, automatic or manual, would take charge. His environment and circumstances elicited a high level of negative emotions. When he found himself in a footbridge-like situation, his automatic system was too active for his manual mode to inhibit. From this, one concludes that Luttrell was not in control of his moral judgment, thereby he should not be held morally responsible for it, and thus should be absolved of his moral regret.

Quickly, we find ourselves in a world where moral behaviors are explained through science. In those cases where automatic processing is involved, there is no moral responsibility,

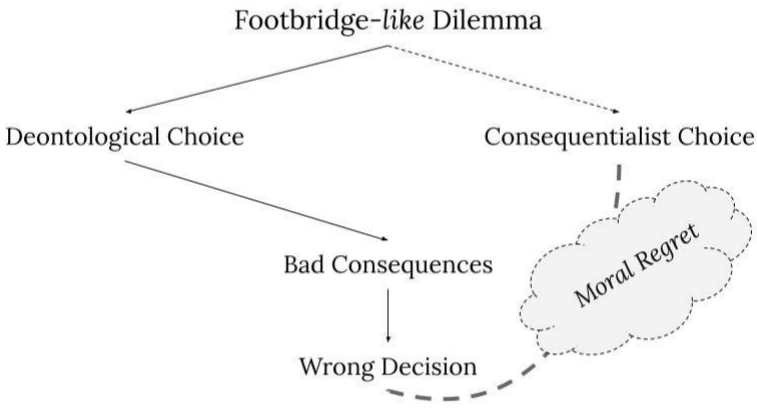
and no warrant for moral regret. We are drones acting upon our unconscious, automatic processes, reacting to environmental pushes and unable to determine our own moral behavior. This approach is unfavorable for two related reasons. First, this idea runs counter to folk psychology – people tend to attribute moral responsibility to moral behaviors. When we are wronged, we hold an aggressor responsible. There are some mitigating circumstances, e.g., mental health issues, that are recognized exceptions to this tendency. However, expanding these exceptions to include behaviors driven by automatic processes means that a wide range of moral behaviors would be deemed beyond our control. With these behaviors outside our control, we find blame and praise to be unwarranted, and thus also find punishment and exaltation unwarranted. In this reading, we let Luttrell off the hook because he is no longer morally responsible for his moral behavior. Our reductio ad absurdum leads us to an unacceptable outcome – a world with no moral responsibility – which we must reject.

Second, in justifying Luttrell's behavior by pointing to what we know about the automatic system, we are robbing him of his moral agency. He is no longer free to choose his action; his deliberation is an illusion. Not only does Luttrell have no free will to choose one act over another, but he also cannot choose what moral laws by which he wishes to be guided. We have moral agency only in behaviors that involve our manual mode, and those behaviors emerging from our automatic system are now non-moral because we have no agency in determining them. I do not believe that this is the lesson to be learned here – that science should be used to decrease our sense of moral agency. Instead, science should be used to increase our moral agency. Learning about our dual-processes should give us information that we can use in our moral deliberations, increasing our moral agency because we are more aware of the previously hidden processes that underlie our moral behaviors.

So, instead of viewing Greene's work as an exculpation of Luttrell's behavior, let's consider a more nuanced view – one where Luttrell is still responsible for his choice, but is also given some relief from his moral regret. Luttrell seems to regret his decision because of its outcome. He views the bad consequences of his action as evidence that he made the wrong decision. Believing he made the wrong decision causes moral regret – he wishes he made the "right" decision. I call this the *naïve view of moral regret*, as it is a familiar reaction to the emotional state that occurs when we make difficult choices. It can be visualized in Figure 3.1a below:

Figure 3.1a

Naïve View of Moral Regret



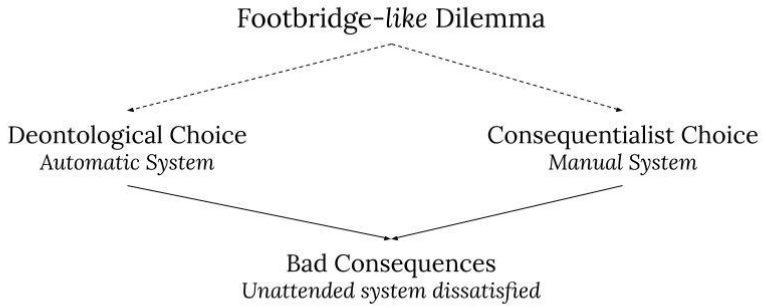
Note: A depiction of Luttrell's moral regret, emerging from the idea that he made the wrong decision when he chose the deontological choice over the consequentialist choice in a footbridge-like moral dilemma.

As an alternative, I propose the following: Luttrell's moral regret is a result of his manual mode being left dissatisfied. To expand, Greene's dual-process theory of moral judgment reveals something important about the cognitive processes underlying our decisions in moral dilemmas.

We now know that we have two separate systems which have evolved to support our moral judgments. Usually, they stay in their respective lanes, helping us make judgments that correspond to our situational needs. However, in some situations, e.g., footbridge-like dilemmas, we feel tension between these systems. Because we feel each system pulling us in a different direction, we are torn in the dilemma, wanting to appease these opposing forces. With either decision we make, we will leave the unappeased system dissatisfied. So, making the deontological choice will leave our manual mode dissatisfied, and making the consequentialist choice will leave our automatic system dissatisfied. The moral regret we feel is caused by the remnants of the unattended system. See Figure 3.1b for a visualization of this process.

Figure 3.1b

A Reinterpretation of Moral Regret



Note: By incorporating Joshua Greene's dual-process theory of moral judgment, we recognize that two evolved systems are giving us competing moral judgments. In a footbridge-like dilemma, we cannot satisfy both. The moral regret we feel is caused by the unattended system being left dissatisfied.

So, instead of viewing Luttrell's moral regret as evidence that he made the "wrong" choice in his dilemma, we should view his regret as a reaction to the remnants of a dissatisfied cognitive processing system.

Luttrell is being pulled toward what his manual mode had promised, i.e., that he would have saved his friends' lives. However, if he had chosen the consequentialist action, he and his friends would have been cold-blooded killers. They may have been charged in US courts or could have faced retaliation from the Taliban. There would have been bad consequences in this situation too. Luttrell would likely feel moral regret and then this regret would serve as evidence that he made the wrong choice – he would be in the very same situation.

I want to defang this idea that Luttrell's moral regrets serve as evidence that he made the wrong choice, or that they're evidence that he did something wrong at all. He would have regrets either way, so this would mean he would make the wrong choice either way, which would mean there was no right choice to make – only two wrong choices. This is a dismal view and one that does nothing but take moral agency away from Luttrell.

Instead, I believe that Greene's research shows us that we are built in such a way that our two systems are triggered in footbridge-like dilemmas. When we serve one of these two systems, it comes at the cost of serving the other. This can leave us with regret – we are still feeling pulled toward the choice we did not make. But this does not track a moral truth about the world; our regret is not evidence that we did the wrong thing. Rather, our moral regret is a longing of the system we chose not to accommodate. Once understood this way, the regret of "I made the wrong choice" becomes "Part of me wishes I had done the other thing." In these dilemmas, there is no escape; we must choose one course of actions over another; but this view allows us to reflect on moral choices of the past and reinterpret our moral regret.

This reinterpretation also contributes to our moral agency. Knowing that in these dilemmas, we will be required to resist one of our evolved systems and that we will feel moral regret because of it, we can make better informed moral decisions. This increases our moral

agency because we are not only choosing which system to satisfy, but also which system to resist, knowing that we will be left with dissonance because of it. Depending on the circumstances, we may not have the time or control to carefully choose one system over the other – it is likely that the automatic system will take charge when we are short on time or control. However, when we do have the opportunity to deliberate on a particular situation, or when we take the time to reflect on our moral actions and principles in a general sense, we can decide better for ourselves which outcome we can best live with.

3.2 Skepticism about Factual Beliefs

The introspective process I have developed calls us to use science to inform our moral deliberations. In this section, we are confronted with an uncomfortable empirical proposal: the factual beliefs we use to inform our moral beliefs may not be as objective as we like to imagine. Instead, we are driven by our *moral* beliefs in such a way that they can shape our *factual* beliefs. Though unsettling, this idea may not be surprising. After all, it seems unlikely that we would have such vast moral disagreement if everyone uses objective facts to form moral beliefs. Even Hume noticed our tendency to be ruled by passion over reason. In the opening of *An Enquiry Concerning the Principles of Morals* (1777), he writes:

> DISPUTES with men, pertinaciously obstinate in their principles, are, of all others, the most irksome… The same blind adherence to their own arguments is to be expected in both; the same contempt of their antagonists; and the same passionate vehemence, in inforcing sophistry and falsehood. And as reasoning is not the source, whence either disputant derives his tenets; it is in vain to expect, that any logic, which speaks not to the affections, will ever engage him to embrace sounder principles (Hume, 1777/1912).

Thus, Hume observes that the power of logic in arguments pales in comparison to the power of affections. Try as we might to persuade a passionate person with our logical reasoning, we may find that they dig in their heels and grow more obstinate.

According to psychological research, at the core of this phenomenon is our desire to minimize cognitive dissonance: when we experience conflict between our beliefs, we seek stabilization. Holding incoherent beliefs causes us distress, and to remedy our discomfort, we adjust those beliefs. Recent views on moral reasoning suggest that the way we achieve this balance is by bringing our *factual* beliefs in line with our *moral* beliefs, rather than the other way around. This phenomenon is the basis for Jonathan Haidt's *Social Intuitionist Model* which, true to Hume, holds that reason is the servant of the passions (Haidt, 2012, p. 58). Although we tend to assume that our moral reasoning is reflective of a private process aimed at finding the truth, it is instead inescapably social and based in intuition. Haidt's model claims that moral reasoning usually follows *after* a judgment has already been made and is heavily influenced by social and cultural influences. Haidt explains that in making moral judgments, the reasoning process is "more like a lawyer defending a client than a judge or scientist seeking truth" (Haidt, 2001, p. 10).

In this section, I discuss two studies which demonstrate the lengths we go to defend our moral judgments: Thomas, Stanford and Sarnecka (2016) and Liu and Ditto (2012). Both studies show that we tend to go as far as to adjust our factual beliefs in order to better support our moral judgment. Thomas et al. refer to this tendency as the *moralized reinforcement of factual beliefs* and show that our moral judgments about parenting affect our assessment of risk to children. Liu and Ditto provide experimental evidence that we shape our factual understanding of the world to fit our moral understanding. In what follows, I briefly describe the findings of each study and

then argue that this empirical knowledge should cause us to adjust our moral deliberations so that we are more critical of our own use of factual information in our moral judgments.

Moralized reinforcement of factual belief

In December 2022, Veronica and Dax Tejera left their two young children sleeping alone in a hotel room in New York City while they dined at a restaurant about a block and a half away from the hotel. They had two cameras streaming footage of the children to their cell phones. While they were out, Dax had a medical emergency – he was intoxicated, choked on food, and was rushed by ambulance to a hospital where he later died. Veronica rode with him to the hospital and called her parents and a friend to ask them to go to the hotel and check on the children. The hotel did not allow the family or friend into the room and instead called the New York Police Department to report that the children were left unattended. Veronica was later charged with child endangerment and currently faces a sentence of up to a year in prison (Margaritoff, 2023).

This story sparked national conversations about child endangerment. In one example, the Huffington Post Parents Facebook page received hundreds of comments about the case. Huffington Post Parenting Reporter Marie Holmes wrote that "most were understanding of Veronica choosing to accompany her husband in the ambulance but unforgiving of the couple leaving the children alone in the first place" (Holmes, 2023). Although this reaction seems natural, or at least predictable, it is also a blatant contradiction. If there is concern about objective risk to the children, why would this risk be emphasized when the parents were at a nearby restaurant watching video of the children sleeping, and then dismissed when the parents are being driven away in an ambulance and one of them is incapacitated? There is an

inconsistency here, where the commentors are making judgments of risk assessments that are mismatched with the objective risk to the children. These assessments are instead well-matched with the commentors' moral judgments rather than the actual danger posed to the children.

In their 2016 paper, "No Child Left Alone," Thomas, Stanford and Sarnecka examine the American parenting norm that has developed in recent decades, which expects children to be under constant direct adult supervision. Like we see in the Tejeras' case, this norm is puzzling in that children who are left alone for even short periods of time are seen as being in more danger than they are in situations like riding in a car, which is objectively more dangerous for them. In fact, one common fear is that an unattended child will be abducted by a stranger. If we look at the data, however, we find that the risk of a teen or child being abducted by a stranger and killed or not returned is around one in 1.4 million each year (Thomas et al., 2016, p. 2). In comparison, there are about 5 deaths per 100,000 children and teens in motor vehicle accidents per year, a rate second only to gunfire, which claims the lives of about 5.5 deaths per 100,000 children and teens per year (Goldstick et al., 2022). Thomas et al. hypothesize that peoples' moral judgments play a role in their risk assessment. In other words, the less morally acceptable a person finds a parent's reason for leaving a child alone, the more danger they think the child is in. Through six experiments, Thomas et al. find support for this hypothesis, leading them to suggest that people overestimate the danger a child is in to better support their moral judgment about why the child was left alone.

The participants across the six studies were given vignettes that described a parent's reason for leaving a child alone, and the circumstances of the act. For example, a 6-year-old may be left alone at a park about a mile from her house for 25 minutes, or a 2.5-year-old may be left home alone, eating a snack and watching *Frozen* for 20 minutes. In all of these vignettes,

Thomas et al. presented different conditions where the parent left the child either unintentionally (e.g., some accident prevented the parent from being with the child), or because the parent had to go to work, or because the parent went to volunteer for charity, or the parent went to relax, or the parent went to engage in an affair. Participants were asked to estimate the risk posed to the child by being left alone in each circumstance.

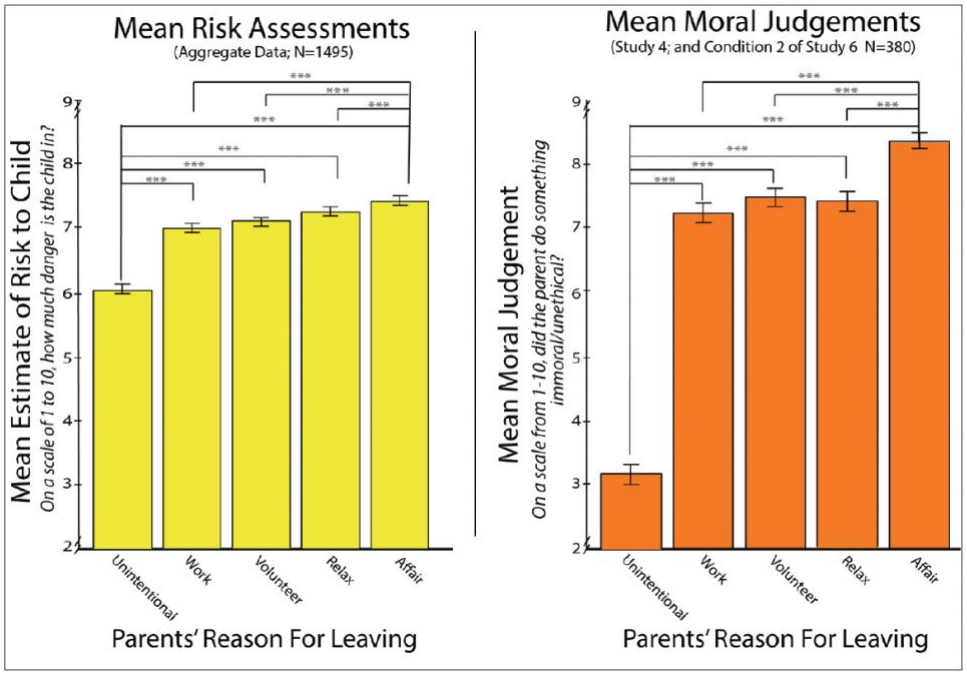
The experiments revealed some surprising outcomes. First, participants regularly assessed the risk to a child who was purposefully left alone as being higher than if the child had been left on accident. In the vignettes shared with the participants, Thomas et al. used identical language in the description of the situation, changing only whether the act was intentional or not. For example, the participant might be told that a 10-month old was asleep in the car in a gym's cool underground parking garage for 15 minutes, but some of these participants would be told this was an accident and others were told it was intentional. Participants considered the child who was left alone intentionally to be in more danger than the child left alone by accident. However, as the researchers observe, it seems more likely that a child left alone on purpose would be in a better position than a child who is unintentionally left alone. In the purposeful case, a parent could set up parameters to create a safer environment for the child – perhaps even positioning cameras to observe and monitor the child, as in the case of the Tejeras.

Second, in comparison to the unintentional condition, participants had increasingly higher estimations of risk for the conditions where the parent goes to work, volunteers, relaxes, or has an affair. The mean risk assessments increased respectively, and can be seen below in Figure 3.2a. When asked to explicitly judge the morality of the parent's actions, the participants' answers were significantly positively correlated with their risk assessments. In other words, the

more immoral the reason for the parent being away, the more danger the participants thought the child was in.

Figure 3.2a

Risk Estimates and Moral Judgments



Note: Participants' responses by moral condition. Left panel: Mean estimate of risk to child by moral condition. Right panel: mean moral judgment by moral condition. Error bars indicate standard error. ***p < .001. (Thomas et al., 2016, p. 7). Used with permission of the University of California Press.

A third interesting finding occurred when Thomas et al. considered whether the absence of a father would have the same effect as the absence of a mother. In most of the experiments, the parent was described as a mother, but in one experiment, the parents were described as fathers instead. The researchers found that there was a similar pattern of judgment across all the situations: a father's affair meant more danger for the child than the father relaxing, which was more dangerous than the father volunteering. However, participants judged fathers who left the child alone to go to work as being about as dangerous as them unintentionally leaving the child.

So, there was a gendered difference between parents, i.e., a mother leaving her child alone to go to work is seen as more dangerous than a father doing the same. Perhaps this finding is not surprising, considering that it corresponds to gender stereotypes about women, especially mothers, being better suited for home-life than the workplace. Extending the researchers' remarks about children likely being objectively safer in scenarios where they are left alone on purpose, though, we might also slow down to examine an inconsistency with this gender norm. If mothers are seen as being more nurturing and attentive than fathers (hence they are better suited for the home than the workplace), wouldn't a corollary be that a mother leaving her child alone would result in greater preparation than in the father's case, thus decreasing the risk to the child? It seems not, and so the moral judgment about mothers cuts only in one direction, to more negatively evaluate a mother who works than a father who works.

In summary, through six studies, Thomas et al. show that people's moral intuitions do in fact affect their risk estimates. When they present two situations with objectively equal risk, the participants estimated higher risk in the situation for which they held moral disapproval. They conclude that "people don't only think that leaving children alone is dangerous and therefore immoral. *They also think it is immoral and therefore dangerous*" (Thomas et al., 2016, p. 12). This adjustment of our risk assessments to fit our moral judgment is coherent with the observations from Hume and Haidt that our reason is subservient to our passions.

Recruiting facts

In a 2012 experiment, Liu and Ditto examine moral dilemmas, investigating how people resolve moral conflict to make a judgment about the right course of action. Moral dilemmas are dilemmas precisely because they require us to weigh various moral considerations against one

another. Liu and Ditto focus on real-world moral dilemmas (e.g., forceful interrogations of terrorist subjects, condom promotion in sex education, capital punishment and embryonic stem cell research), noting that while people tend to have interpersonal conflict with others over these topics, they are typically not internally conflicted. For example, a person who believes that capital punishment is morally justifiable is likely to also believe that it is effective as a deterrent.

So, why is it that we disagree with others about the right course of action, but we avoid internal conflict about our own views? This may not be a surprise considering the effect we saw moralized belief have on factual belief in the parenting case. We tend to reduce internal conflict by aligning our factual beliefs in accordance with our moral beliefs. In their paper, Liu and Ditto present three important findings which give us a deeper understanding of how our descriptive cost-benefit analysis is influenced by our prescriptive moral beliefs. The researchers begin with a familiar finding: the more immoral we find an act, the more harmful we predict its consequences. Next, they discover that as moral conviction grows stronger, the coherence between moral and factual beliefs increases. Finally, Liu and Ditto find that reading morality-based essays on a topic can change one's factual beliefs, even if the essays do not contain any fact-based information. I examine each finding below and then present a case for how Liu and Ditto's study can help us in our moral decision making.

First, Liu and Ditto find that people align their moral evaluations of an act with beliefs about its consequences (Liu & Ditto, 2012, p. 318). This is important in one sense because it supports the other findings and theories discussed earlier. We see that this phenomenon is not only related to parenting judgments but occurs more broadly, and this gives more credence to the theory of cognitive dissonance and Haidt's social intuitionist model. Another, novel discover Liu

and Ditto made in their study is that people aligned their moral and factual beliefs in both artificial moral dilemmas and real-world moral dilemmas.

In examining participants' reactions to artificial moral dilemmas, Liu and Ditto used the *footbridge* Trolley Problem example, where the participant must decide whether to push a large stranger onto the trolley tracks in order to save a group of workers who are in the trolley's path of destruction. The researchers asked the participants whether any number of lives saved would justify pushing the stranger over the bridge. Eighty percent of the participants responded in a by saying that there is no trade-off that would justify the act. These participants, when compared to the others who would make some trade-off, believed that pushing the stranger off the bridge was less likely to actually save lives, and more likely to cause pain to the stranger.

Might it be that such a bizarre situation like the Trolley Problem confuses participants or puts them in an unfamiliar mindset where they are forced to choose an action they have never contemplated taking? To determine whether the artificial nature of the Trolley Problem influences how people respond, Liu and Ditto gave real-world moral dilemmas to participants. The artificial moral dilemma findings were replicated: participants' moral beliefs about forceful interrogations, condom promotion, capital punishment and stem cell research, significantly predicted their beliefs about the issue's costs and benefits. Again, Liu and Ditto show that there is a strong association between the moral evaluations of an act and its positive and negative consequences.

When examining participants' evaluations of real-world issues, Liu and Ditto found that the level of moral conviction is associated with beliefs about the likelihood of outcomes. The more participants believed an action was moral, the more they believed it would produce beneficial consequences and the less they believed it would have undesirable costs. Likewise, the

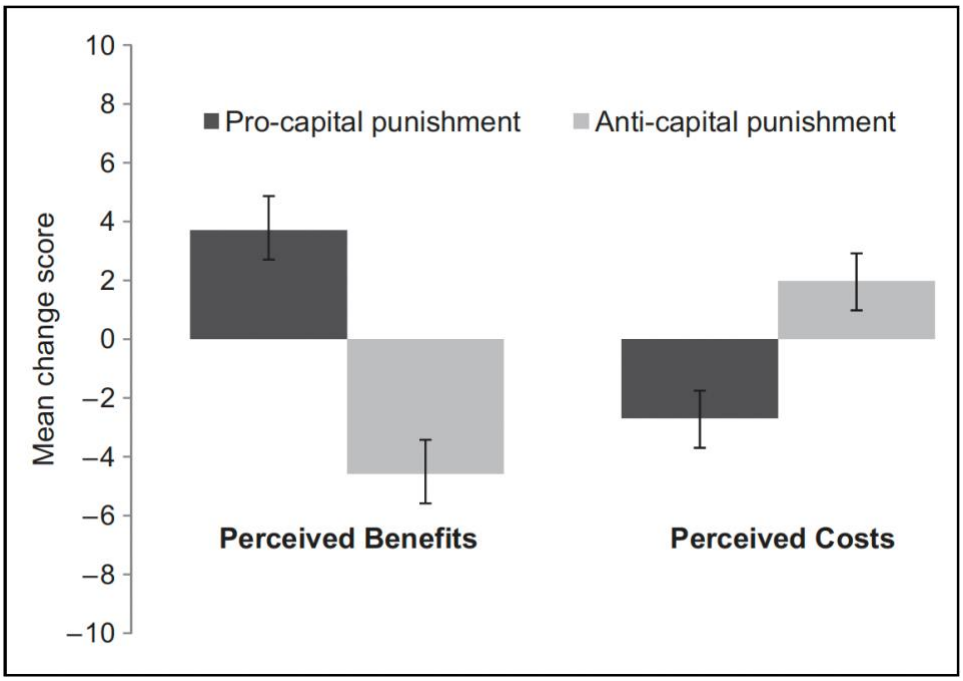
more participants believed an action was immoral, the less they believed it would produce beneficial consequences and the more they believed it would have undesirable costs. For example, in the case of condom promotion in sex education, the more a participant believed that promoting condom use was morally wrong even if it prevented pregnancy and STDs, the less they believed condoms were effective at such prevention and the more they believed that promoting condom use would encourage teens to have sex (Liu & Ditto, 2012, p. 318). Furthermore, participants with stronger convictions also perceived themselves as being more informed about the issues. So, a participant with the above views about condom promotion would like also feel highly informed about the issue. Liu and Ditto note research (Goodwin & Darley, 2008; Skitka et al., 2005; Wright et al., 2008) that has shown that positions held with moral conviction are often experienced as objective, self-evident truths (Liu & Ditto, 2012, 319). A coherent picture emerges, then, of a person with strong moral convictions, corresponding cost/benefit predictions, and confidence that they have the relevant facts. The relationships between these markings, however, are only correlational; they do not yet show the hypothesis that drives Liu and Ditto's study: that the moral convictions are influencing the factual beliefs.

To investigate a causal relation between moral convictions and factual beliefs, Liu and Ditto designed an experiment where participants would first answer questions about the morality, benefits and costs for the four real-world issues. Second, the participants were randomly assigned to read an essay with moral arguments either for or against capital punishment. Importantly, the essays made no mention of factual information like costs and benefits; the arguments were purely moral in nature. Finally, the participants re-answered the same capital punishment questions they were asked before reading the essay. If the participants' factual beliefs are influenced by moral beliefs, as Liu and Ditto predicted, then there should be some

movement in their beliefs about costs and benefits that align with the moral/immoral view of capital punishment that they read. In fact, this is exactly what Liu and Ditto find. Although it is a small effect, the researchers found a statistically significant change in participants' factual beliefs in the direction predicted by the moral argument they read. For example, a participant who read an essay in favor of capital punishment would express stronger beliefs that capital punishment works as a deterrent and weaker beliefs that it is carried out against an innocent person. Figure 3.2b illustrates that participants who read the pro-capital punishment essays had a positive change in perceived benefits, i.e., that capital punishment works as a deterrent, and a negative change in perceived costs, i.e., that capital punishment is used against innocent people. Similarly, participants who read the anti-capital punishment essay had a negative change in the perceived benefits and positive change in the perceived costs of capital punishment. Liu and Ditto note that the changes seen in this experiment were small, even though they were significant. However, this is not surprising given the minimal moral influence of the study on the participants. They only read a short essay, approximately 520 words. A more involved experiment may elicit stronger effects.

Figure 3.2b

Experimental Changes in Factual Belief



Note: Mean change in perceived benefits and costs by essay condition. Positive change represents believing capital punishment to be more beneficial and costly after the essay (Liu & Ditto, 2012, p. 320). Used with permission of Sage Publishing.

Through their experiments, Liu and Ditto find that factual beliefs correlate with moral beliefs in both artificial and real-world dilemmas, that this correlation becomes stronger as moral conviction grows, and finally that factual beliefs can be influenced by exposure to moral arguments. From this, they conclude that "people shape their descriptive understanding of the world to fit their prescriptive understanding of it" (Liu & Ditto, 2012, p. 321). This should prompt us to accept that the naïve view of factual beliefs informing our moral beliefs, is at least sometimes misleading. Rather, our moral beliefs influence the factual beliefs that we form. They influence how we assess risk and how we predict costs and benefits. Furthermore, the stronger a person's moral convictions, the more extreme their predictions of the facts *and* the more they think they are informed about the topic. The passions are not only in control of the reasons; they have also convinced us that the reasons are in control.

An honest deliberation

Both the Thomas et al. and the Liu and Ditto studies support the Social Intuitionist Model, showing that we go to great lengths to justify our moral evaluations, even revising our factual beliefs to make them fit with our moral intuitions. The impact this has on society should not be ignored. Parents are being arrested for leaving their children alone *even in low-risk situations*. They are being charged with endangering their children, but if there is no (or very little) danger, it seems they are actually being punished for violating norms.

My project calls for an informed introspection, where we use empirical knowledge about our moral processes to make the moral decisions we want to make. Knowing that our factual beliefs are influenced by our moral beliefs should prompt us to invite some skepticism into our appeal to facts in moral deliberations. When we recognize that we are supporting a particular moral judgment or intuition by appealing to factual beliefs, we should subject those factual beliefs to heightened critical scrutiny. We should demand concrete, objective evidence that these beliefs are true before we rely on them. Or, to take the lead from Liu and Ditto, we may opt to expose ourselves to moral arguments from positions with which we disagree. Doing so may serve to check and balance our factual beliefs, where we take multiple factors into consideration: the moral arguments, the available factual information, as well as a barometer for our own moral convictions.

In our deliberative process, these changes can make us more honest with ourselves (and others) by focusing on our moral principles. We can come to see our factual beliefs as instantiations of our intuitions. For example, when I deliberate about capital punishment, I could think to myself, “The idea of the state ending the life of one of its citizens violates my deeply

held moral principles. This is confirmed to me when I am compelled by the alarming number of executed persons who have later been found to be innocent. People who disagree with me usually focus on data that claims that the death penalty acts as a deterrent to capital crimes. I am more moved by the thought of an innocent person being executed than I am by the thought of a guilty person walking free. Maybe the other person's focus on crime deterrence can tell me something about their values." We often justify our moral beliefs by pointing to our factual beliefs. Whatever value that phenomenon has to us, it is one we do not actually fulfill. Not only do our moral judgments disconnect from our moral principles, but they act in a circular way, justifying the facts we use to justify our moral judgments. Our moral deliberations are meant to provide us reflection and clarity about our moral judgments and ridding them of the illusion that our factual beliefs are morally neutral will allow us to better employ them. Ultimately, this will bring about a more honest deliberative process, where we can bring our moral decisions in harmony with our moral principles.

3.3 An Informed Introspection

These two examples demonstrate how science can and should be used to make concrete changes to our moral deliberation. In the first case, our understanding of the dual-process theory of moral judgment gives us reason to reinterpret the moral guilt we're left with when our two evolved systems conflict. We can choose for ourselves which system we wish to serve and which we will resist, with the understanding that we will still live with the remnants of the unappeased system. This will offer us some relief from moral regret, which we will no longer see as evidence that we made the wrong choice. The second case calls our attention to our factual beliefs,

showing that they are often influenced by our moral judgments. Thus, when we deliberate about moral situations, we should be careful with using our factual beliefs as grounds for our actions – they may not be as objective as we would hope.

In the next chapter, I provide another example of how knowing the science behind our moral processes can help us improve our moral deliberations. In this case, we find that our moral intuitions are inundated with group biases. Although parsing our social world into categories is not in itself a bad thing, and can be quite useful in many ways, this is a process of which we should be aware. Our moral intuitions are susceptible to these biases, causing us to unconsciously form attitudes about others that we consciously reject. This incoherence can be ameliorated by bringing our attention to it: becoming skeptical of our moral intuitions can help us correct for the ways they lead us astray from our moral principles.

We are no longer blind to our cognitive biases when we bring them to the surface. We recognize that our automatic cognitive processes are efficient but not very flexible. They help us make quick judgments and predictions, but they may cause us undue suffering, in the case of moral regret, and may not serve the moral principles we hold, in the case of factual beliefs. Training ourselves to see these cognitive processes for what they are can help us reframe the role that these automatic tendencies play in our moral deliberation.

Importantly, the message here is not simply, *People are biased*. Rather, the informed introspection is critically self-reflective. The message it conveys is, *Science tells me that my moral behaviors are at times influenced by cognitive biases. If I care to act in ways that align with my moral principles, I should be aware of these biases and correct for them. Doing this will contribute to my moral agency, allowing me to act in ways I think are right*. The informed introspection will enhance our agency by supporting our ability to self-govern. Not only is this a

view that aligns with Korsgaard's views on moral agency and self-government, but in the final chapter, I connect the informed introspection to Dewey's ideas about morality as well as Rawls's view of reflective equilibrium. My goal is that this will show broad philosophical significance for my view.

CHAPTER 4: Skepticism of Moral Intuition

4.0 Moral Intuitions

Humans have a deeply embedded inclination to group ourselves and each other into social categories. This ability allows us to navigate our complex social world: understanding group membership allows us to reason about how others think, what they believe, how they will act and how they will interact with one another (Liberman et al., 2017, p. 1). This inclination is not simply a categorization tool, though. Entrenched within our social organization are preferences: we tend to like people who are similar to us. We also tend to like people who conform to their group. As Gordon Allport explains, everywhere on earth we find groups of people who stick to their own ingroup. They mate, eat, play, and worship with their own kind. This automatic cohesion is convenient – we do not need to exert the energy it takes to adjust to new languages, foods, beliefs, and so on (Allport, 1979, p. 17).

Chapter 3 revealed how relevant our automatic processing is to our moral judgments. Our automatic system is rife with moral intuitions – strong, stable and immediate moral beliefs (Sinnott-Armstrong et al., 2010, p. 246). These intuitions influence our moral behaviors, yet they are not regulated by our conscious reasoning. While the tendency to prefer our own kind is automatic and ubiquitous, it can also result in prejudice and discrimination toward outsiders. For those with egalitarian values, this means that maintaining fairness and equality ultimately meets resistance. Such tension lies at the heart of this chapter.

In this chapter, I highlight research that explores the development of social categorization, looking at how our moral intuitions (i.e., the attitudes we unconsciously develop

toward others) are influenced by implicit bias. Implicit bias is the idea that we can act on the basis of prejudice and stereotypes without intending to do so (Brownstein, 2015). This means that we may treat others according to stereotypes of their respective social groups without any awareness that we harbor these stereotypes. There has been much psychological research done on implicit bias and how it is present throughout the spectrum of our social lives: education, healthcare, law enforcement, employment, and so forth.[24] Importantly, my target here will be implicit bias rather than explicit bias. My project is meant to help moral agents gain more control over their moral deliberations so that they may act in a way that better aligns with their own moral principles. Illuminating how our implicit biases influence our moral intuitions serves this purpose. Correcting for explicit biases, conversely, would entail calling on agents to change the moral principles they endorse, which is not in the purview of my project.

I begin by discussing how little is needed to activate our social categorization – even something as simple as a sticker can prompt children to think in terms of ingroup and outgroup membership, i.e., us versus them. Next, I look at how we develop expectations that others will conform to their respective group norms and punish them when they fail to do so. Then, I examine how our attitudes emerge from these social categorizations and expectations, as well as how they predict our treatment of others. Finally, I use a recent Supreme Court case to exemplify how understanding the influences over our moral intuitions can help us build in a healthy dose of skepticism to our moral deliberations in a way that will give us more agency over our moral decisions.

[24] See Jost et al. (2009) for a review of the empirical research on implicit bias.

4.1 Ingroup and Outgroup Membership

One primal way humans organize our social world is by sorting others into ingroup and outgroup categories. That is, we find others to be like us or not like us. This tendency is considered by psychologists to be effortless and automatic (Turner et al., 1979). For example, when we step into a room full of people, we will likely form automatic social categories of gender, race, and age without any conscious effort (Brewer, 1988; Fiske & Neuberg, 1990). We can even make multiple categorizations simultaneously, though our behavior tends to follow the most salient group divisions. So, if we are at a conference for women in STEM, gender will probably be a more salient grouping than race – we'd notice the gender groups in the room more than we'd notice different race groups – though we would likely categorize for both.

Psychologists can elicit ingroup and outgroup categorizations in experiment participants even under minimal conditions, i.e., when the difference between us and them is a matter of some novel social division. In this minimal group paradigm, researchers create novel groups with which participants have no experience – for example, assigning individuals to groups by giving them matching stickers with arbitrary colors. This allows the researchers to control for variation in individual experiences, opinions, and motives (Tajfel et al., 1971; Simon & Gutsell, 2020, p. 2). Not only are minimal groupings effective in prompting us/them thinking, but minimal groups can even be more salient than race groupings, directing behavior in a way that prioritizes novel ingroup/outgroup membership over race membership (Van Bavel & Cunningham, 2009). In other words, if we were given stickers to mark our group membership in a room full of people, we would notice whether other people are in or out of our sticker-group more than we'd notice their race group.

Minimal grouping conditions show us just how easily we mark others as part of us or them. This is important because our attitudes toward our ingroup and outgroup follow closely: we prefer those who are like us over others. This preference is visible in our evaluations and our behaviors – we think better of our ingroup members and we treat them better too, even when our group membership is totally arbitrary (Baron & Dunham, 2015; Baron & Banaji, 2006; Tajfel et al., 1971; Van Bavel & Cunningham, 2009). So, if we prefer those who are like us, and the thing that makes them "like" us is arbitrary, it follows that we can prefer others for arbitrary reasons. Emphasizing that social categorization is an automatic process, we are unconsciously developing attitudes toward others that can be totally arbitrary – not based on core values that we consciously involve in our moral deliberations. It seems that our moral intuitions are being formed independently of our moral principles. If we want to bring these into alignment, we should begin by investigating how our social categorization works and then build this knowledge into our moral deliberations.

Baron and Dunham (2015)

To further examine minimal grouping conditions, I look to a study by two developmental psychologists, Andrew Scott Baron and Yarrow Dunham, in 2015. Baron and Dunham (2015) explored how group membership affects children's attitudes and how children learn about social groups. To set up minimal grouping conditions, they presented children, 6 to 8 years old, with cartoon illustrations that vaguely resembled humans – they had faces, bodies and limbs. Two groups were distinguished, "Lups" and "Nifs," and members of each group had the same skin color, either red or purple. The children were given stickers that identified them as a member of

one of the two groups; the stickers had the name of the group and an illustration of a member from their group with the appropriate skin color, red or purple.

In the first experiment, Baron and Dunham gave the children eight ambiguous situations where one Lup and one Nif interacted. Half of the situations included a negative social behavior, like knocking someone over, and the other half included a positive social behavior, like helping a friend at school. For each situation, the children were asked attribution questions like "Who knocked the person over?" or "Who helped their friend with their schoolwork?" The experimenters found that children were more likely to generalize negative behaviors to outgroup members than ingroup members. For example, a child who had been given a Lup sticker would be more likely to say that a Nif had knocked someone down than that a Lup had knocked someone down. Furthermore, the children were more likely to generalize positive behaviors to ingroup members than outgroup members. So, a child who had been given a Nif sticker would be more likely to say a Nif helped someone with their homework than that a Luf had helped someone with their homework.

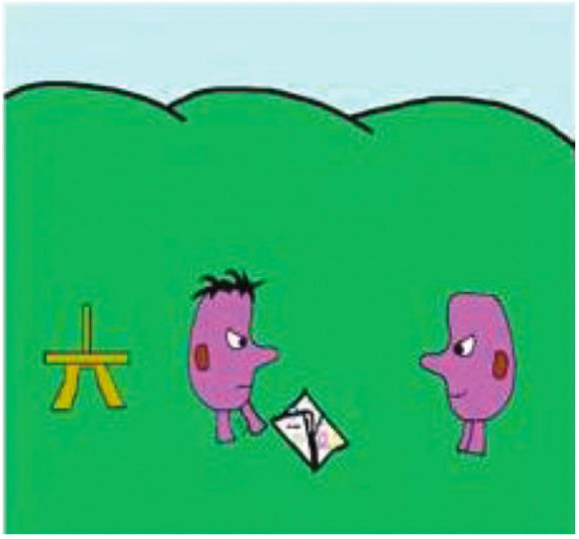
Next, the experimenters measured the preferences of the children. The participants were shown a Lup and a Nif side-by-side and explicitly asked which one they liked more. The children reported a stronger preference for their ingroup members; they chose ingroup members over outgroup members about 57% of the time (Baron & Dunham, 2015, p. 6). Experiment 1 shows how quickly ingroup positivity emerges. The children developed this bias based only on shared group membership carried by a label. They were not, for example, given information about competition between groups or cooperation within groups; there was no mention of expectations for future interactions between groups. Thus, the experimenters showed that an ingroup bias comes "for free," without prior knowledge or surrounding expectations (Baron &

Dunham, 2015, p. 6). Knowing nothing beyond whether an individual was part of their group was enough for children to judge whether the individual acted positively or negatively in an ambiguous situation. It was also enough to develop a preference for those in their group over those outside their group.

In two subsequent experiments, Baron and Dunham explored what would happen if the situation was not ambiguous. In Experiment 2, children were given objective evidence that individuals from one group acted antisocially, e.g., one individual tearing up the artwork of another individual (see Figure 4.1). The children were assigned to one of the two groups and then shown information characterizing either their ingroup or the outgroup in a negative way. For example, a child assigned to the Lups group would either be shown information about two Lups tearing up a Nif's artwork, or two Nifs tearing up a Lup's artwork. One-third of the children were assigned to the transgressors' or actors' group (e.g., a Lup seeing two Lups tear the Nif's artwork), one-third were assigned to the victim's or recipient's group (e.g., a Nif seeing two Lups tear the Nif's artwork), and one-third were not assigned to either group, making them third-party observers.

Figure 4.1

Groups in Minimal Conditions



Note: Sample illustration from Baron and Dunham (2015), Experiment 2. Two members from one group are observed intentionally damaging someone's painting (Baron & Dunham, 2015, p. 8) Used with permission of the Taylor and Francis Group.

The researchers asked attribution questions, like they did in Experiment 1. This time, though, they also asked the children to infer about new behaviors. They measured the children's judgments about how likely new members of both groups would be to engage in behaviors not previously depicted in the story (Baron & Dunham, 2015, p. 9). The results of Experiment 2 showed that children who were members of the actors' group were less negative in their judgments than children from the other two conditions. For example, a Lup child who saw Lups act badly was less likely to make negative attributions toward the Lups than would a child who was part of the Nifs or a child who was a third-party observer. From this, Baron and Dunham suggest that group membership partially protects us from making negative inductions about the behavior of our ingroup members (Baron & Dunham, 2015, p. 11). In other words, we start off with a bias that prevents us from recognizing the bad behavior of our ingroup members.

As in the first experiment, Baron and Dunham asked the children about their preferences. In comparing preferences before and after observing the Experiment 2 situation, children from

the recipient's group and children from the third-party group both showed sharp declines in their preferences for the actor's group. Children in the actors' group, however, were not as affected in their preference. In fact, these children had a nonsignificant increase in preference after seeing members of their group acting badly. So, for example, a Lup child who saw Lups act badly would probably not be swayed away from her pre-test preference, and her preference for the Lups may increase by a small (nonsignificant) amount. From this, Baron and Dunham suggest that belonging to a group protects children from reliably internalizing negative information about their own group (Baron & Dunham, 2015, p. 12). To summarize, knowing that members of a group have acted badly toward another individual will have a significant effect on those from the recipient's group or a third party – their preference for the group will sharply decline. However, this knowledge probably will not have any impact on the actors' own group members.

In Experiment 3, Baron and Dunham explored what would happen if the children from the actors' group were given more balanced information. In Experiment 1, the children received only ambiguous information, and in Experiment 2, they received only negative information. So, in Experiment 3, the researchers gave the children information about the actors engaging in two negative and two positive behaviors. The children were still divided into three groups: those from the actors' group, those from the recipient's group, and third-party observers.

Again, the researchers first asked the children attribution questions (e.g., asking who engaged in the behavior of the story) and inference questions (e.g., asking about the likelihood of a new group member engaging in new behaviors). The results of Experiment 3 did not show any group-related differences in terms of judging the actors' group positively or negatively. That is, if a child saw Lups acting toward a Nif positively twice and negatively twice, it didn't matter

whether the child was also a Lup or a Nif or a third-party observer – all groups tended to judge the actors' group similarly.

Baron and Dunham note two possibilities for this result. One possibility is that without as much negative information, there just was not enough information to get the biases to an observable magnitude. Another possibility is that because there were positive and negative scenarios, the information was more complex and more difficult to track and internalize (Baron & Dunham, 2015, p. 16). Interestingly, even though the children were given an equal number of positive and negative scenarios, they tended to judge the actors negatively overall. The children predicted that the actors' group would have more negative behavior (than the recipients) in the future and less positive behavior (than the recipients) in the future. Baron and Dunham suggest that negative information is weighed more heavily than positive information.[25]

Like Experiment 2, the children in Experiment 3 differed in their preferences. Children from the recipient's group grew in their dislike (Mean = -13%) of the actors' group. Third-party observer children also grew in their dislike (Mean = -8%) of the actor's group, but not as much as those from the recipient's group. Again, there was a non-significant increase in the actors' group preference (Mean = +2%). So, when the children saw members of one group act positively and negatively toward a member of another group, they sharply increased their dislike if they were a member of the recipient's group, they somewhat increased their dislike if they were a third-party observer, and they stayed around the same in their preference – or even increased it a little bit – if they were part of the actor's group.

[25] This is consistent with other research on the influence of negative information. See Baumeister et al. (2001) for a review.

From Baron and Dunham's study, we find that group membership is incredibly important for attitudes. They showed that when children see someone as an ingroup member, even based on minimal conditions, they will likely prefer the individual to someone they see as an outgroup member. Furthermore, they showed that in an ambiguous situation where it is not clear whether someone acted badly, children are likely to think that an ingroup member acted positively and an outgroup member acted negatively. Importantly, when a situation is not ambiguous, and children are given clear information that an ingroup member acted negatively, they still prefer the ingroup member. So, the children are able to generalize negative behavior to the individual and predict that the individual will act badly, but they still prefer the individual over outgroup members.

Researchers in developmental psychology continue to investigate the early origins of our ingroup preferences. In as early as the first year of life, infants show a tendency to like similar individuals and dislike dissimilar individuals. Hamlin, Mahajan, Liberman and Wynn (2013) studied the preferences of 9- and 14-month infants, finding that they prefer individuals who treat similar others well and treat dissimilar others poorly (Hamlin et al., 2013, p. 589). In their experiment, Hamlin et al. determined whether the infants preferred graham crackers or green beans. Then they showed the infants two rabbit puppets: a similar rabbit who liked the same snack as the infant and a dissimilar rabbit who liked the other snack. The researchers then showed either the similar or dissimilar rabbit playing with a ball and introduced two dog puppets. One dog, the helper, would help the rabbit by returning the ball to it. The other dog, the harmer, would take the ball and run away with it. Then they presented the helper and harmer puppets to the infants to see which one they reached out for, marking their preference. Overwhelmingly, the infants preferred the dog that helped a similar rabbit and the dog that harmed a dissimilar rabbit. Hamlin et al. are careful to note that this study did not demonstrate that the infants created social

categories of "green-bean lovers" or "graham-cracker lovers" or even ingroups of "individuals like me" or outgroups of "individuals not like me." However, these early preferences are likely related to the similarity preferences shown by children and adults. Hamlin et al. conclude that we have an inborn or early-developing propensity to like those whom we recognize as similar to ourselves and dislike those who differ from us (Hamlin et al., 2013, p. 593).

In Section 4.3, we return to ingroup preferences, investigating them as attitudes that predict behavior. As another example of how our group psychology develops, I next turn to how children and adults evaluate individuals who resist their respective group norms. We find that children expect individuals to conform to the behavior or their own groups, whether that behavior is positive or negative. In short, children prefer conformists to non-conformists.

4.2 Descriptive-to-Prescriptive Tendency

Recall from Chapter 2 the attempts and failures of Kitcher and Greene to avoid the naturalistic fallacy. Even though each was acutely aware that it was an illicit move to gather normative consequences out of their descriptive accounts, the lure was too strong. Ultimately, they each made the leap from is to ought as they conflated what is natural with what is good. As Hume observed, we are easily tempted to push the is/ought boundary. Whereas Chapter 2 provided an analysis of why this conflation is philosophically problematic, this current section examines our psychological tendency to conflate these two concepts in the first place.

In a recent series of investigations, psychologist Steven O. Roberts and his colleagues have examined the cognitive underpinnings of our tendency to infer an "ought" from an "is." This descriptive-to-prescriptive tendency begins in children as young as four years old and

declines with age although it is still present in adults. The descriptive-to-prescriptive tendency happens as we start to observe behavioral regularities in groups. We notice that members of social groups may tend to act in similar ways (e.g., that boys tend to like sports). Roberts et al. have found that children and adults do not only notice these descriptive regularities, but they also use them to infer prescriptive norms (e.g., that boys should like sports) and negatively evaluate those who do not conform to their group's regularities (e.g., a boy who doesn't like sports). In other words, we tend to use our observations about how members of a group do act to form judgments about how they should act.

To illuminate the full picture of the descriptive-to-prescriptive tendency, I discuss two of Roberts's studies. I use the first (Roberts, Gelman and Ho, 2017) to explain the tendency itself and how it can be seen even under minimal conditions where individuals are behaving in activities that have a neutral valence (e.g., eating berries). In the second study (Roberts, Ho and Gelman, 2019), I discuss what the researchers found when they examined how the tendency works under conditions of positive or negative behavior (e.g., helping a ladybug or hurting a ladybug). Together, these findings and others show us that conformity plays a large role in our evaluations of others.

Roberts, Gelman and Ho (2017)

In their paper, "So it is, so it shall be: Group regularities license children's prescriptive judgments" Roberts, Gelman and Ho ask: When do descriptive regularities become prescriptive norms? In order to assess the descriptive-to-prescriptive tendency in minimal conditions, Robert et al. designed three studies with important differences from previous research. First, they asked about two novel groups, Glerks and Hibbles, rather than using social categories that the

participants were already familiar with (e.g. race or gender). The only difference between Glerks and Hibbles was their clothing pattern (green stripes or orange triangles). Second, Roberts et al. did not ask about behaviors with moral footings. Settling whether this tendency arises in morally irrelevant conditions can help determine the minimal conditions under which it occurs. If morally relevant behaviors were considered, the participants' responses could be confounded by their previous cultural and social commitments. Third, the participants were not part of either group. This is important because if they were, they would likely have bias against out-group members and preference for in-group members that are prior to their evaluations of the individuals' behaviors (Baron & Dunham, 2015).

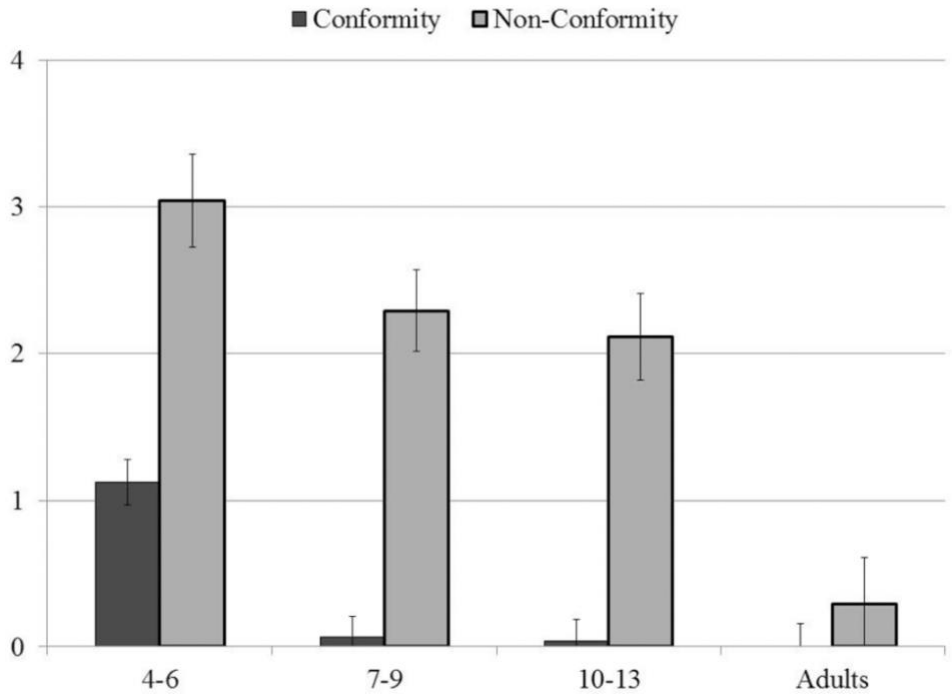
Roberts et al. studied participants in four age groups: 4-6 years old, 7-9 years old, 10-13 years old and adults. In Experiment 1, the researchers told the participants about Hibbles and Glerks, attributing properties to these novel groups. For example, the researchers would show both groups, which were on different sides of the screen. Then they would point to different colored berries, saying "Hibbles eat these kinds of berries [pointing to one color berry] and Glerks eat these kinds of berries [pointing to the other color berry]." Other behavior domains included types of toys the groups played with, the languages they spoke, and the music they listened to. Then, for each of eight trials, they would show a conforming or non-conforming individual from one of the groups. The researchers then asked the participants whether the conforming or non-conforming individual's behavior was "okay" or "not ok." When the participants answered "not ok" the experimenters followed up with a scale to determine if the participants thought the behavior was "a little bad, pretty bad, or very, very bad" (Roberts et al., 2017, p. 5). Additionally, the participants were asked for open-ended explanations for their

responses, which were broadly coded by the researchers into themes about group, individuality, norms and similarity.

The results of Study 1 showed that all age groups most often approved of conforming behavior. All the child age groups were significantly more disapproving of non-conformity than conformity. The youngest children were the most disapproving of non-conformity, and this effect declined with age. In fact, only one of the twenty-four adults in the study disapproved of non-conformity. The open responses showed that the youngest children explained their disapproval of non-conformity by citing normative rules. Older children were more likely to cite group membership to explain why they disapproved of non-conformity. When individuals, including adults, approved of non-conformity, they typically gave individual-based explanations, e.g. the individual's wishes or desires.

Figure 4.2a

Disapproval of Non-Conformity I



Note: Results from Roberts, Gelman and Ho (2017), Study 1. Mean frequency of disapproval of conformity across age groups. Scores could range from 0 to 4. Error bars depict standard errors (Roberts, Gelman and Ho, 2017, p. 22). Note that all age groups of participants disapproved of non-conformity, with the strength of disapproval waning with age. Used with permission of John Wiley & Sons publications.

Roberts et al. write that the results of Study 1 suggest that children view group-based regularities as having prescriptive force (Roberts et al., 2017, p. 7). Even with novel groups, of which they were not a part, and with innocuous behaviors like eating berries, the children tended to think that the regularities of groups should be followed by their members.

At this point, however, Roberts et al. questioned whether their results were an artifact of assumed competition or cooperation dynamics. For example, did the children negatively evaluate the non-conforming individuals because they thought they would jeopardize within-group coordination, leading to vulnerability in between-group competition? In other words, were they worried that a Hibble who ate orange berries like the Glerks would jeopardize the Hibble-coordination, making it weaker when in competition with the Glerks? So Roberts et al. designed Study 2 to account for these possible confounds.

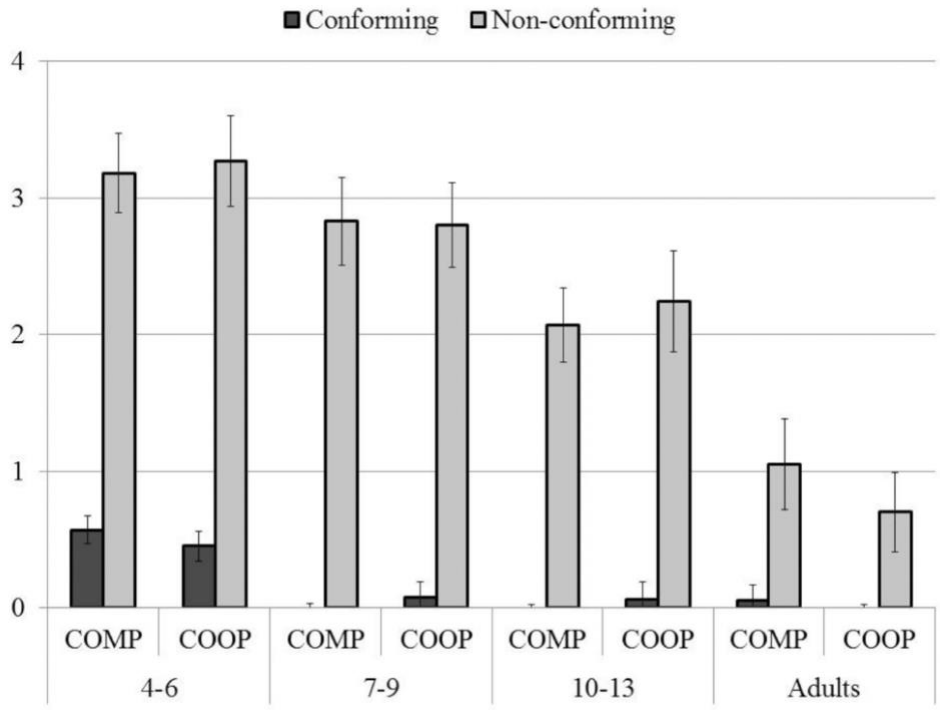
In Study 2, participants were assigned to either a competition or cooperation condition. Here, the Hibbles and Glerks were said to be building towers out of blocks. In the competition condition, the groups were competing for a prize and there were limited resources (blocks). In the cooperation condition, the groups were building a tower together and there were enough blocks for everyone.

The results from this study showed that all age groups were more disapproving of non-conformity than conformity (Roberts et al., 2017, p. 9). The children were more disapproving than adults, with the youngest children being the most disapproving. Importantly, there was no main effect for the condition, meaning that the results held across the cooperation and

competition conditions. In fact, Roberts et al. explain that they "obtained identical patterns in both the competition and cooperation conditions" (Roberts et al., 2017, p. 11). This means that in both conditions, non-conformity was evaluated more negatively than conformity, and that there was a decline of disapproval with age. From this, Roberts et al. speculate that the descriptive-to-prescriptive tendency may be so important to human psychology that it emerges early in development and holds robustly across intergroup contexts (Roberts et al., 2017, p. 11).

Figure 4.2b

Disapproval of Non-Conformity II



Note: Results from Roberts, Gelman and Ho (2017), Study 2. Mean frequency of disapproval of conformity and non-conformity across age groups and conditions. Scores could range from 0 to 4 (Roberts, Gelman and Ho, 2017, p. 23). Note that all age groups of participants disapproved of non-conformity. Used with permission of John Wiley & Sons publications.

Study 3 aimed to distinguish whether the descriptive-to-prescriptive tendency is particular to group-contexts or if it also emerges with descriptions of individual regularities.

Only children were recruited for this study since they showed the strongest descriptive-to-prescriptive tendency in the previous two studies. Roberts et al. stripped the instructions of any reference to groups but kept the same clothing patterns as previously established so that there would only be a visual marker of group membership. The children were shown two individuals wearing different clothing patterns and on different sides of the screen. They were told that each individual has a different property, e.g. the individual wearing green stripes listens to one kind of music while the individual wearing orange triangles listens to another kind of music. Then the children were shown another individual whose clothing matched one of the initial two individuals. The third individual either conformed or did not conform to the pattern established by their respective shirt-matching individual. For example, the third individual is wearing green stripes but non-conforming because she listens to the music that the individual wearing orange triangles listens to.

Here, Roberts et al. found that disapproval was only marginally higher for non-conformity than for conformity (Roberts et al., 2017, p. 12). There were, however, three important differences between Study 1 and Study 3. First, the rate of disapproval for non-conformity was higher in Study 1. Second, the rated negativity (from 0 to 4, reflecting a little bad, pretty bad, or very, very bad) in Study 1 was greater than it was in Study 3. Third, the children in Study 1 were more likely to use norm-based explanations when asked about their negative evaluations than were the children in Study 3. From these results, Roberts et al. suggest that the negative evaluations children have of non-conforming individuals stems from group regularities rather than regularities in general.

Roberts et al. conclude by writing that “these studies provided converging data showing that with regard to third-person, unfamiliar, and morally neutral groups, the link between what is

and what should be is powerful in childhood" (Roberts et al., 2017, p. 14). Together, these studies forge important ground in understanding the minimal conditions under which the descriptive-to-prescriptive tendency operates.

Roberts, Ho and Gelman (2019)

Another study by Roberts, Ho and Gelman (2019) investigates whether the descriptive-to-prescriptive tendency holds when group norms are uncommon or when they involve a positive or negative behavior. The researchers were looking to test the limits of the tendency: would it hold even when the group norm was uncommon, like raising a foot to ask a question or drinking orange juice out of a bowl? Would the tendency hold even when the group norm was a negative behavior, like punching a person? Roberts et al. designed two experiments to explore these questions. They included three age-groups of participants: 4- to 6- year-olds, 7- to 9-year-olds, and adults.

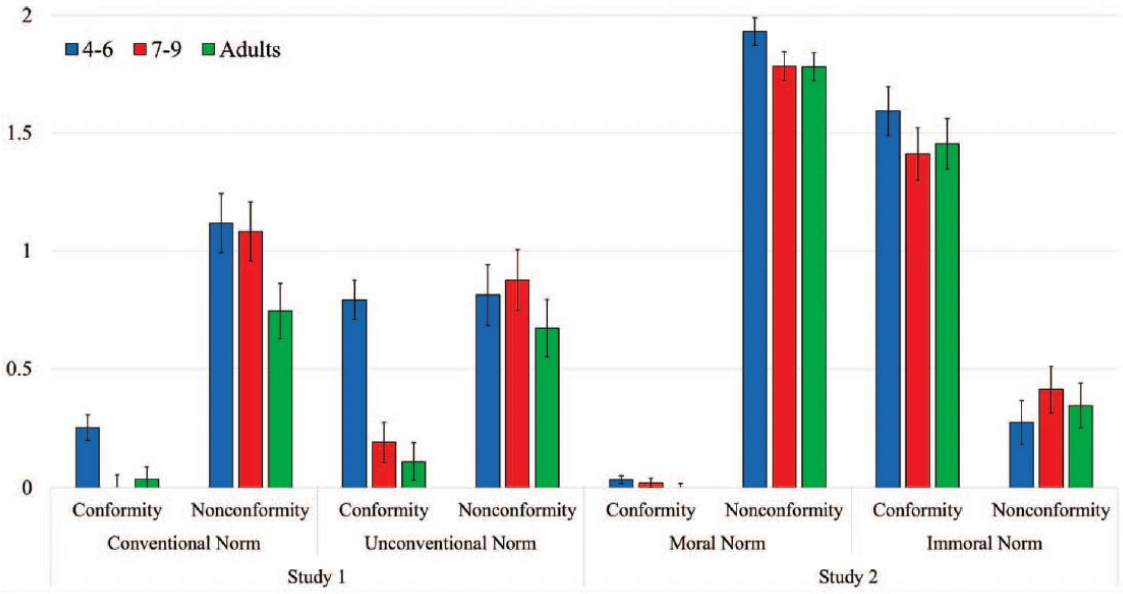
In the first experiment, the researchers presented two novel groups, Hibbles and Glerks, and common and uncommon norms as group regularities. Examples of common norms were: drinking out of a cup, raising a hand to ask a question, eating cake at birthdays, or using a leash to walk a dog. Examples of uncommon norms were: drinking out of a bowl, raising a foot to ask a question, eating beans at birthdays, or using a leash to walk a cat. Individual Hibbles or Glerks were then shown to either conform or not conform to their group's norms. Roberts et al. found that the descriptive-to-prescriptive tendency holds despite the prevalence of the norm. Adults and children were more disapproving of non-conforming individuals than of conforming individuals and that it did not matter whether the group regularity was a common or uncommon norm.

In the second experiment, the researchers asked participants to evaluate individuals (again, from the novel groups, Hibbles and Glerks) who were either conforming or not conforming to their group's positive or negative behaviors. For example, the participants were shown a Hibble, who unlike other Hibbles, made babies cry; or they would be shown a Glerk who, like other Glerks made babies smile. The researchers found that participants were much less likely to follow the descriptive-to-prescriptive tendency when evaluating positive and negative behaviors. That is, children and adults disapproved of individuals when they committed negative behaviors, regardless of whether they were conforming to their group norms (Roberts et al., 2019, p. 383). Interestingly, though, there was still an effect of this tendency. First, even though all age groups of participants disapproved of individuals who behaved badly across the board, the level of disapproval depended on whether or not the individual conformed with their group norm. For example, the participants disapproved more of individuals who didn't conform to positive group norms (e.g., someone who, unlike their group, made babies cry) than they disapproved of individuals who conformed to negative group norms (e.g., someone who, like their group, makes babies cry). So, being from a group that makes babies cry gives someone a little leniency when they make babies cry – they would be judged harsher if they were from a group that makes babies smile. In other words, non-conformity seems to make a negative behavior more negative.

Similarly, all age groups disapproved more of individuals who did not conform to negative group norms than they disapproved of individuals who conformed to positive group norms. For example, an individual who made babies smile unlike their group was judged more negatively than someone made babies smile like their group. So, non-conformity seems to make a positive behavior less positive.

Figure 4.2c

Disapproval of Non-Conformity III



Note: Results from Roberts, Ho and Gelman (2019). Mean frequency of disapproval rates of each age group across study and behavior type. Scores could range from 0 to 2. Bars depict standard error (Roberts, Ho and Gelman, 2019, p. 379). Note the descriptive-to-prescriptive tendency regardless of prevalence or valence of the group norm. Used with permission of the American Psychological Association.

The significance of this study rests in its finding that children do not blindly follow a descriptive-to-prescriptive tendency. Roberts et al. suggest about children that "their beliefs about what is most common or good can override how they expect group members to behave" (Roberts et al., 2018, p. 384). There are competing principles at play. Roberts et al. explain that children and adults seem to find uncommon behaviors more acceptable if they're what a group is known for, and they find common behaviors less acceptable if they're not what a group is known for. Children and adults also seem to find it more acceptable to act in negative ways if it's what a group is known for, and less acceptable to act in positive ways if it's not what a group is known for.

Like Roberts, Gelman and Ho (2017), these findings from Roberts, Ho and Gelman (2019) contribute to a growing picture of the strength and prevalence of the descriptive-to-prescriptive tendency. Under minimal conditions, children and adults showed a tendency to expect individuals to conform to their group norms, and to evaluate them negatively when they fail to conform. The bottom line is that whether an individual is seen as conforming to their group plays a significant role in how we judge them.

One final note about the descriptive-to-prescriptive tendency is that it declines with age – the effect is typically present but not as strong in adults. Roberts et al. (2017) take up this question, flagging it for future research. One possible explanation they suggest is that adults require more than minimal conditions to elicit the tendency – they may need more social context or more serious moral violations to activate the tendency. Another possible explanation is that we learn to suppress this tendency as we grow older. This latter suggestion fits well with the research on implicit and explicit attitudes, which shows that with age, we express less biased attitudes toward social groups like race and gender, but our implicit bias remains fairly constant. In the next section, I examine this research as I look at attitudes and behavior.

4.3 Attitudes and Behavior

Psychologists distinguish between implicit and explicit attitudes. As mentioned in the beginning of this chapter, we are unaware of our implicit biases. Explicit biases, however, are the attitudes that we self-report. For example, the children in Baron and Dunham's (2015) study reported their explicit attitudes when they were asked which individual they preferred, the ingroup member or outgroup member. Likewise, the children and adults in Roberts et al.'s

studies were also asked to explicitly evaluate conforming and non-conforming individuals. To measure implicit attitudes, researchers cannot depend on participants' self-reporting. Instead, they turn to tools like the Implicit Association Test (IAT).

The IAT is a well-known tool for measuring implicit bias in adults.[26] It works by measuring how strongly a participant associates concepts, e.g., groups, with good/bad evaluations or stereotypes. The underlying assumption is that when the participant automatically associates the concept with an evaluation or stereotype, it will be easier for them to respond and thus their reaction time will be lower. This measurement is considered to reflect a person's implicit attitudes because it is measuring the ease/difficulty with which a person matches a group to an evaluation or stereotype rather than explicitly asking them to self-report their preferences.

Much of the research done on implicit and explicit biases involves social groups, and specifically our attitudes toward race groups and gender groups. So, although the previous work in this chapter involved a purposeful exclusion of social context, like race, to examine explicit preferences under minimal conditions, the work I discuss in this section will directly assess implicit and explicit attitudes regarding race. There are two important findings from this research that I will highlight: (1) explicit bias tends to decline with age while implicit bias tends to persist, and (2) it is our implicit, not explicit, attitudes that predict our treatment of others in socially sensitive situations.

Development of Implicit and Explicit Attitudes

In a 2006 study, Andrew Scott Baron and another researcher, Mahzarin R. Banaji, examined the difference between the development of implicit and explicit attitudes. They

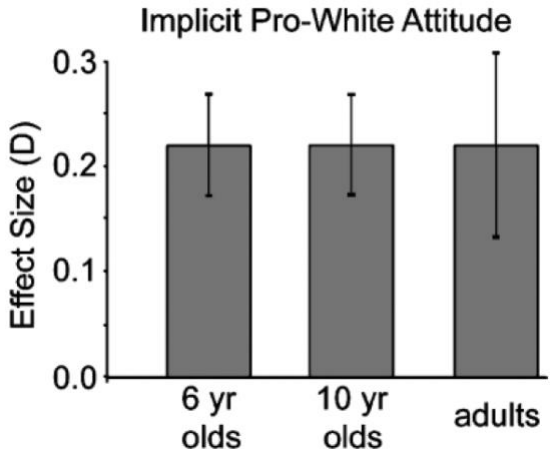
[26] See Jost et al. (2009) for a review of the IAT and the history of objections against it.

measured race attitudes in White American 6-year-olds, 10-year-olds, and adults. To measure implicit attitudes, the researchers developed a child-friendly version of the IAT.

Baron and Banaji (2006) began with 6-year-olds because children are known to reason about race in a similar way to adults from around 5 years old. By this age, they have begun to essentialize racial kinds, believing one's race to be fixed and immutable (Hirschfeld, 1996, 2001). In this study, Baron and Banaji found a significant average IAT effect, D = 0.22, indicating a pro-White/anti-Black response bias. Importantly, the magnitude of implicit race bias was the same among all age groups: the 6-year-olds, 10-year-olds, and adults (see Figure 4.3a). So, as far as implicit attitudes are concerned, the participants overall showed ingroup preferences that were stable from childhood to adulthood.

Figure 4.3a

Implicit Race Preference

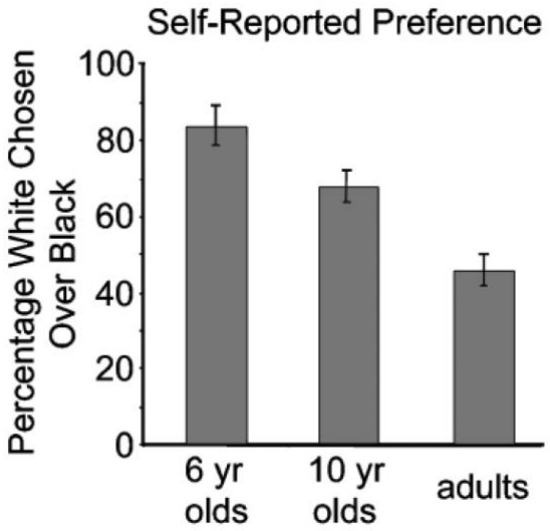


Note: Results from Baron and Banaji (2006). Implicit race preference in three age groups. A positive value of D indicates a preference for Whites relative to Blacks (Baron & Banaji, 2006, p. 55). Note that there is no magnitude of difference between the implicit pro-White attitudes of White Americans from three age-groups. Used with permission of the Association for Psychological Science.

A different story emerged, however, when the participants were explicitly asked whether they preferred White or Black children. When Baron and Banaji asked the participants to self-report their preferences, the 6-year-olds chose a White child over a Black child 84% of the time. This preference dropped in the 10-year-olds, who chose a White child over a Black child 68% of the time. In adults, this preference disappeared, with adults selecting a White child over a Black child only 46% of the time (see Figure 4.3b; Baron & Banaji, 2006, p. 55-56). So, the youngest age group showed a strong preference for members of their ingroup, but this preference subsides by about age 10, and then disappears in adulthood.

Figure 4.3b

Explicit Race Preference



Note: Results from Baron and Banaji (2006). Explicit race preference in the three age groups (Baron & Banaji 2006, p. 55). Note that explicit preferences for White children over Black children declines and disappears from childhood to adulthood. Used with permission of the Association for Psychological Science.

To summarize, Baron and Banaji showed an asymmetry in the development of implicit and explicit attitudes. In their study, implicit attitudes favoring the ingroup remained stable

across development. Explicit attitudes, however, began strongly in favor of the ingroup but became more egalitarian over development. The researchers suggest that this shows a divergence between implicit and explicit attitudes around 10 years old, with explicit attitudes becoming more sensitive to the societal demand for unbiased race-based evaluation (Baron & Banaji, 2006, p. 57). In other words, as children develop, they begin to align their explicit attitudes with the accepted values of their society.

Adults often deny that they have ingroup preferences, especially when it comes to race. As the Baron and Banaji (2006) study shows, we self-report egalitarian preferences. In fact, the concept of racial colorblindness is widely endorsed as a way to manage diversity and intergroup relations (Apfelbaum et al., 2012). In professing themselves as colorblind, adults commit to the belief that racial group membership should not matter, or be used or acknowledged, in many settings. Proponents of the concept claim that this mindset will prevent discrimination. Evan P. Apfelbaum, a social psychologist, has studied the practice and implications of color blindness in interpersonal, educational, organizational, legal, and societal domain. In a review, Apeflbaum and fellow researchers found that the practice of color blindness is far from a cure for discrimination. Instead, it can hinder race relations and create more problems than it solves (Apfelbaum et al., 2012). For example, in an experiment on color blindness in children, Apfelbaum, Pauker, Sommers, and Ambady (2010) found that children exposed to a color-blind mindset were less likely to identify overt instances of racial discrimination and less likely to describe the discrimination in a serious way than students who were exposed to a value-diversity mindset. In another experiment, Apfelbaum, Sommers, and Norton (2008) found a number of negative consequences when White participants employed a colorblind strategy in a social context. Not acknowledging race when in social settings where race was relevant led to White

participants displaying more negative nonverbal behavior, suffering cognitive impairment due to the inhibitory control diverted needed to avoid acknowledging race, and being seen as more racially prejudiced by Black participants (Apfelbaum et al., 2008).

When we claim not to have an ingroup racial bias, we are professing our egalitarian norm. We are committing to view people of different races equally, without preference for our own racial group. However, the research shows that this is often in tension with our implicit attitudes. The development of our implicit and explicit attitudes diverges around age 10. So, even though we are likely to self-report egalitarian preferences as adults, we are also likely to retain an implicit preference for our ingroup over our outgroup. Because our implicit attitudes are automatic and unconscious, we are unaware of the role they play in our treatment of others. In the next section, I discuss how these two attitudes influence our behaviors.

Predicting Behavior

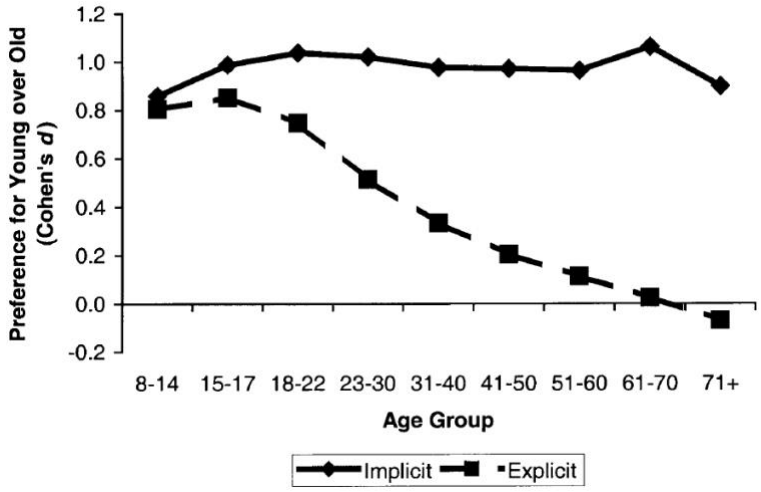
So, can we choose to follow our explicit attitudes or are we doomed to blindly follow our implicit attitudes? It turns out that the answer is a bit nuanced. Social psychologists have found that overall, our implicit and explicit preferences tell similar stories – they tend to have a positive correlation (Nosek et al., 2002). We are not two entirely different people living in one body; our implicit and explicit attitudes often match up. However, the strength with which they correspond differs across domains.

For example, in a 2002 study, Nosek, Banaji and Greenwald found a small correlation between respondents' implicit and explicit attitudes toward age, as seen in Figure 4.3c. In the youngest respondents, implicit and explicit attitudes were very similar. However, as the age of the respondent increased, their implicit preference of young over old held fairly steady even

though they were far less likely to self-report such a preference. In other words, young people are likely to be both implicitly and explicitly biased against old people, and the older they get, they stay implicitly biased but they are much less likely to self-report any bias.

Figure 4.3c

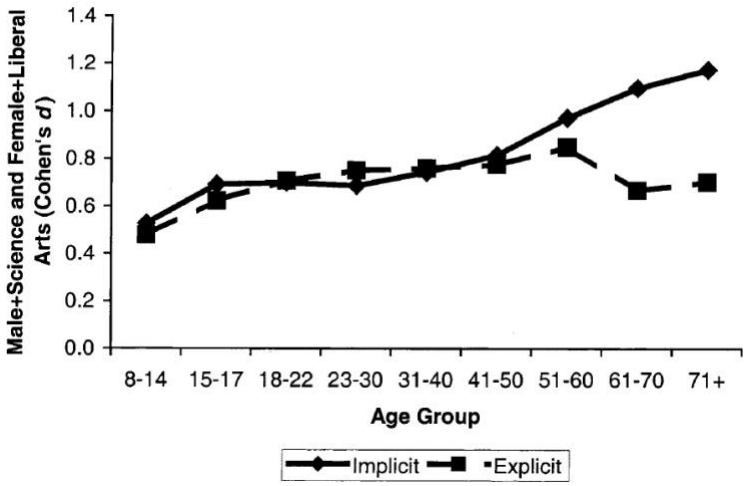
Preference for Young Over Old



Note: Results from Nosek, Banaji and Greenwald (2002). Implicit and explicit attitudes toward young versus old are reported as a function of respondent age. (Nosek et al., 2002, pp. 107-8). Note the implicit and explicit preference toward young versus old. Used with permission of the Educational Publishing Foundation.

Figure 4.3d

Implicit and Explicit Stereotypes Between Gender and Field of Study



Note: Results from Nosek, Banaji and Greenwald (2002). Implicit and explicit stereotypes linking male with science and female with liberal arts by respondent age (men, n = 19,906; women, n = 36,547). Positive Cohen's ds reflect male with science and female with liberal arts associations; negative values reflect male with liberal arts and female with science associations (Nosek et al., 2002, p. 109). Used with permission of the Educational Publishing Foundation.

Conversely, when measuring respondents' implicit and explicit stereotypes linking males to science and females to liberal arts, respondents' explicit and implicit attitudes strongly corresponded until the oldest age groups, as seen in Figure 4.3d. So, respondents under about age 50 were likely to express explicit bias that roughly matched their implicit bias. In the older age groups, the respondents' implicit bias grew stronger and they were less likely to self-report any bias.

So, why is it that respondents' implicit and explicit attitudes toward age diverged while their implicit and explicit attitudes toward gender/field of study stereotypes mostly converged? Nosek et al. suggest that there are complex cultural constraints that influence the convergence or divergence of our implicit and explicit attitudes. Importantly, they note that we should not think of either of these attitudes as the "real" attitude of the person – we should resist the notion that a person's "true" attitude is their implicit bias, or even that it's what they profess as their explicit

preferences. Instead, we should understand that all social groups hold implicit biases. In our social world, this can mean different things. For example, when it comes to explicit ingroup race preferences, both White and Black respondents showed a strong ingroup preference: White respondents preferred White people over Black people and Black respondents preferred Black people over White people. However, when it comes to implicit ingroup race preferences, both White and Black respondents preferred White people over Black people. In other words, Black respondents lacked ingroup implicit preferences, and even preferred their outgroup. This is consistent with other studies examining preferences of Black participants and other nondominant racial and ethnic groups, e.g., Latino Americans (Dunham et al., 2013; Dunham et al., 2014; Newheiser & Olson, 2012; Baron, 2015; Baron & Dunham 2015). Nosek et al. suggest that for members of these nondominant groups, their implicit attitudes reveal the influence of negative attitudes held by the culture toward the groups (Nosek et al., 2002, p. 112). So, implicit and explicit attitudes diverge and converge depending on cultural factors like how socially acceptable a stereotype is, or how a society marginalizes a group.

When implicit and attitudes diverge, Nosek et al. ask us not to consider one the "real" attitude of the person – and I agree, we do not want to reduce a person to their implicit or explicit attitudes, especially considering the cultural and social factors that are likely at play. However, if we want to use this information in a way that will help us better control our own moral decisions, we should seek understand how our attitudes influence behavior. We should ask whether implicit or explicit attitudes are more reliable predictors of our behaviors.

In a 2009 meta-analysis, Anthony. G. Greenwald and colleagues reviewed 122 research reports that used the IAT to predict behavioral, judgment, and physiological measures. The review found that both IAT implicit measures and explicit measures, e.g., self-reporting, were

predictive. However, when it came to socially sensitive topics, the predictive validity of self-reporting was impaired (Greenwald et al., 2009, p. 17). This is consistent with research that shows prosocial behavior toward Black people is negatively predicted by implicit prejudice (Ashburn-Nardo et al., 2003; McConnell & Leibold, 2001; Rudman & Lee, 2002) and that discriminating against female job applicants is predicted by implicit stereotypes (Rudman & Glick, 2001; Rudman, 2004, pp. 132-133).

So, generally speaking, both implicit and explicit attitudes can predict behaviors, judgments and physiological measures. However, implicit attitudes are more reliable than explicit attitudes in predicting responses that are socially sensitive in nature. In these situations, the predictive validity of self-report measures was "remarkably low" (Greenwald et al., 2009, p. 32). A cohesive story is emerging here. From Nosek et al., we learned that social and cultural factors can influence implicit and explicit attitudes differently. From Baron and Banaji (2006), we learned that explicit (but not implicit) attitudes become more responsive to social values like egalitarianism as we age. Now, from Greenwald et al. (2009), we learn that when a topic is socially sensitive, implicit attitudes are better predictors of behaviors than explicit attitudes. The preferences we profess seem to carry less weight in our behaviors than the preferences we unconsciously carry. What, then, should we do with this information? Next, I argue that we should use this information about our attitudes to add a healthy dose of skepticism about our moral intuitions to our moral deliberations.

4.4 Moral Deliberations

We unknowingly make decisions all the time that may be influenced by our implicit biases. Who do we sit next to in the doctor's waiting room? Which servers do we tip generously when dining out? Which job applicants do we recommend? Who do we approve for a mortgage? When do we feel a person poses a potential threat or is dangerous? Knowing what the empirical research says, it is easy to acknowledge that our implicit attitudes affect our decision making from the seemingly mundane to the critically substantive situations. So, how should we build this knowledge into our decision-making?

To illustrate how we should implement our newfound scientific knowledge, I turn to a recent U.S. Supreme Court case, Kennedy v. Bremerton School District. I will not offer a legal analysis, nor a directly moral one. Rather, I show that if that if the petitioner involved, Joseph Kennedy, were to consider the empirical research about our moral cognition, his moral deliberation may be changed in a way that would give him more moral agency over his actions. As a result, Mr. Kennedy will be better able to live up to the moral values by which he wishes to be governed.

In June 2022, the Supreme Court Justices ruled 6-3 that a high school football assistant coach, Mr. Kennedy, should not have been suspended by his school district for praying on the field after games. For eight years, Mr. Kennedy routinely prayed on the 50-yard line after his team played. Early on, students began asking him what he was doing. He would tell them he was praying, and some students asked to join him in prayer. He did not turn them away. Eventually, an opposing team's coach made a comment to the school's principal, calling it "pretty cool" that Mr. Kennedy was allowed to pray on the field (Liptak, 2022). That comment caught the attention of the administration, who instructed him not to pray if it interfered with his duties or involved students. Eventually, Mr. Kennedy was suspended, and a school official recommended against

renewing his contract for the following season. The Supreme Court ruling resulted in Mr. Kennedy being reinstated as an assistant coach of the Bremerton High School football team.

In an interview with the New York Times, Mr. Kennedy explained how he saw his role as a coach who sometimes leads student athletes in prayer. When parents would object, Mr. Kennedy took no action against their students: he did not see himself as expressing favoritism to people based on whether they prayed with him (Tavernise, 2022). However, in the Supreme Court hearing, the justices took up the question of whether students could feel coerced into prayer. Justice Brett Kavanaugh asked, "What about the player who thinks, 'If I don't participate in this, I won't start next week'?... Every player's trying to get on the good side of the coach… I don't know how to deal with that, frankly" (Liptak, 2022). As I outline below, Justice Kavanaugh's concern is supported by the empirical research we have reviewed. If Mr. Kennedy were to involve this research in his moral deliberation, he may reconsider his assertion that he would not show favoritism toward students who prayed with him. He would find himself needing to grapple with Justice Kavanaugh's question.

To work our way through Mr. Kennedy's moral deliberation – and what my view can contribute to it – I examine it in five steps: (1) the salient group membership conditions, (2) concerns about conformity, (3) Mr. Kennedy's explicit attitudes, (4) Mr. Kennedy's implicit attitudes, and (5) Mr. Kennedy's informed introspection. This step-by-step analysis serves as a model for how we should deliberate about similar situations.

(1) *Salient Group Membership Conditions*

There is a salient ingroup/outgroup division for Mr. Kennedy: those who join him in prayer and those who do not. According to the research we have reviewed, people have an

automatic and effortless tendency to group others into us and them categories, even under the most minimal circumstances. Recall that even though we can process multiple categorizations at once, our behavior follows the grouping with the most salience (Van Bavel & Cunningham, 2009). Considering that Mr. Kennedy was engaging in a deeply personal act of faith and that he was conscious that his act drew controversy, it is easy to infer that the salient categorization in this circumstance was whether or not students prayed with him. In fact, considering what we have learned about how grouping works, if students from the opponent's side came to pray with him, their membership in his prayer ingroup would probably influence his attitude toward them more than would their membership in the opposing football team.

(2) *Concerns About Conformity*

The descriptive-to-prescriptive tendency tells us that we expect people to conform to their group's norms, and we think of them negatively if they do not. Mr. Kennedy's prayer events became attractions after each football game. As more and more student players joined him in prayer, it could be argued that praying became a group norm. In Roberts's studies, group norms were communicated by saying something like, "Hibbles eat these kinds of berries [pointing] and Glerks eat these kinds of berries [pointing]" (Roberts et al., 2017, p. 4). If students heard comments like "Bremerton football players pray after games," this may be enough to present prayer as a group norm. Recall that in Figure 4.2c, individuals who did not conform to a moral group norm were judged more harshly than anyone else. The prayers were likely seen as a moral group norm by many of those from the school community, and especially by Mr. Kennedy. Justice Kavanaugh's concern, then, becomes palpable – who would want to be seen as a nonconformist?

(3) *Mr. Kennedy's Explicit Attitudes*

Mr. Kennedy self-reports egalitarian attitudes toward his players, regardless of whether they engaged in prayer with him. Recall that Baron and Banaji (2006) found that in the case of race group preferences, explicit ingroup preferences start to decline in children at around ten years old, and adults self-report vigorously egalitarian attitudes. Nosek et al. found that cultural constraints play a significant role in shaping our explicit attitudes and Greenwald et al. found that in socially sensitive situations, explicit attitudes are poor predictors of behavior. The empirical research points to our explicit attitudes being heavily influenced by social and cultural factors. Given that Mr. Kennedy was aware his prayers were controversial, and presumably aware of laws separating church and state, it is no wonder his self-reported attitudes were egalitarian in nature.

Importantly, as Nosek et al. (2002) remind us, we should not think of implicit or explicit attitudes as our "true" attitude. Both are meaningful, and in many aspects of life each can strongly predict our behavior. We have no empirical reason to doubt that Mr. Kennedy's professed egalitarianism is insincere. He may very well feel that he shows no favoritism to students who pray with him. In fact, for my purposes here, I will absolutely take Mr. Kennedy's claim at face value: I accept that Mr. Kennedy values fair treatment of his student players. The problem, however, is not his explicit attitudes. The problem is that Mr. Kennedy thinks his explicit attitudes are all that matter. We know that is not the case. In this socially sensitive situation, Mr. Kennedy's implicit attitudes are more important to consider.

(4) *Mr. Kennedy's Implicit Attitudes*

As we have seen, measuring implicit attitudes requires tools like the IAT. Since we cannot ask people to self-report their implicit attitudes, we must rely on proxies, e.g., comparing reaction times to paired words. We have no such data from Mr. Kennedy, so it is impossible to say what his personal implicit biases may be. However, we have examined what the literature in social and developmental psychology tells us about factors that influence our implicit attitudes. From Baron and Banaji (2006), we learned that implicit ingroup preferences stay stagnant through development, unlike explicit ingroup preferences. Also unlike our self-reported attitudes, they do not seem to be susceptible to cultural and social factors. So, in a very reasonable and understandable way, Mr. Kennedy likely sees students who pray with him as his ingroup and likely implicitly prefers them to those who do not pray with him. Positing the existence of such biases is no condemnation on Mr. Kennedy; this is an automatic process over which it is assumed Mr. Kennedy has no control.

Let us briefly consider implicit racial bias, as this is the subject of most implicit bias research. Some researchers believe that we automatically categorize race – it happens in milliseconds and is difficult to suppress (Ito & Urland, 2003; Park & Rothbart, 1982; Hewstone et al., 1991; Stangor et al., 1992; Van Bavel & Cunningham, 2009, p. 322). Furthermore, as we learned through the research on color blindness, suppressing our racial biases can backfire (Apfelbaum et al., 2008). So, are we just left to sit with our implicit biases? If Mr. Kennedy does have implicit biases for those who pray with him and against those who do not, is he helpless in the matter? Must we succumb to the unconscious process that favors us toward those "like" us in whatever arbitrary way becomes salient? Our moral intuitions seem to be out of our control: acknowledging implicit bias seems to rob us of moral agency.

(5) *Mr. Kennedy's Informed Introspection*

My goal here is not to rob Mr. Kennedy of his moral agency, but rather to give him the opportunity to recalculate his moral decision. In using scientific knowledge about how our categorization of ingroup and outgroup members can affect our attitudes toward others and thus our behaviors toward them, Mr. Kennedy will gain control his moral deliberation. What was once a simple profession of his explicit attitude can now become an informed introspection. Mr. Kennedy, and all of us, since we are all likely to be in his shoes at many times throughout our lives, should reflect on what science has told us about our moral intuitions.

Mr. Kennedy's original moral deliberation involved only his explicit attitudes. He self-reports that he does not show favoritism, which indicates he values fairness. In his deliberation, Mr. Kennedy likely thought something of the sort, "I would never take away an opportunity from a player just because they don't join me in prayer." When, undoubtedly, situations came up where players were given or denied opportunities on the field, Mr. Kennedy probably had explicit justifications for these decisions which did not include whether or not the player prayed with him. Many of us have done the same, thinking: "I sat next to this person in the waiting room because there was an open chair close to the door" or "I tipped low because the server forgot about our table for 20 minutes and I was really thirsty" or "I hired this applicant because he went to a really good college" or "This couple has steady jobs so they'll be able to handle this mortgage" or "The other side of the street is better lit, so I'll be safer over there while this person walks by." Our retroactive moral reasoning process, and how our factual beliefs are influenced by our moral beliefs, was exposed in Chapter 3. Here, I stress that our moral deliberations may involve self-reports of many attitudes, beliefs and values. In line with Nosek et al. (2002), we should not think of one of these as "true" and the other "false." The important move we should

make is not to view our self-reported attitudes as the end of the story. Instead, we should listen to what the empirical research about implicit bias is telling us: we have automatically carve the world into groups based on who is like us and who is not. Furthermore, this unconscious process reliably influences our attitudes, biasing us in favor of those like us. In socially sensitive situations, it is these implicit biases that are more predictive of our behavior. Our explicit attitudes should mark the beginning of our moral deliberations, not the end.

Involving knowledge of our implicit biases in our moral deliberations requires that we acknowledge the empirical work on them. We should be open to what scientific processes have discovered about how we categorize others and how that process influences our attitudes and behavior. This helps us avoid problems that arise when we deny our implicit bias, which were apparent in the example of color blindness. Researchers have suggested various effective ways in which we can combat our implicit biases. For example, Van Bavel and Cunningham (2009) found that creating social groupings unrelated to race, even in the most minimal ways, may shift automatic racial biases. Although this essentially means trading in one ingroup bias for another, it allows us the opportunity to have more control over what biases we find acceptable or unacceptable. I would take sticker-based bias over race-based bias any day.

Mr. Kennedy may choose to group students in a way unrelated to their prayer participation before he makes important team decisions, like by how many yards they ran in the last game. Admittedly, it is likely Mr. Kennedy already does this in some way while he is making coaching decisions – we already granted that he probably has other reasoning in mind like, "I'll start this player because he made great tackles in the last game, not because he prayed with me." What I am asking Mr. Kennedy to do is to build skepticism of his own moral intuitions into his moral deliberation. He should guard himself against implicit biases he knows he is likely

to have. He should reason, “Science has taught me that I’m likely to favor the players who pray with me. I value fairness, and I do not want to show favoritism based on whether students pray with me. So, I will take measures that scientific research shows can help me shift my implicit biases in a way that does not violate my value of fairness.” This process is consciously directed by Mr. Kennedy and he has the discretion to insert his own values. The science does not validate nor repudiate his moral core. Instead, it boosts his moral agency so that he is choosing for himself the principles by which he wants to be guided.

This chapter has examined how our automatic categorization of our social environment leads us to form biases for some people and against others. Crucially, our preferences are not a result of our conscious reasoning processes, nor are they borne out of our deeply held moral values. Instead, as we have learned, we develop preferences based on factors like whether someone is in our ingroup or whether they conform to their group norms. When we use minimal grouping experimental practices, we find that something as simple as sharing the same sticker as others can set off these preferences. This is important because it shows that our preferences can develop without any moral basis. In other words, I can simultaneously (a) believe that the type of sticker a person wears is morally irrelevant, (b) develop implicit biases in favor of those who wear my shared sticker and against those who wear others, and (c) behave in ways that are influenced by these implicit biases. I am thus acting in a prejudicial way that was initiated by a factor that I admittedly find morally irrelevant.

The solution I aimed to provide to this quandary is that science can help us build better cohesion between our moral intuitions and moral behaviors. By understanding how our attitudes and behaviors are influenced by our social categorization, we gain opportunity to curtail their power. We can build into our moral deliberative process a skepticism of our moral intuitions.

Instead of relying on our explicit attitudes to explain our behaviors, we will understand that our implicit attitudes are very powerful, especially in socially sensitive situations. So, when we find ourselves in socially sensitive situations, we can be careful about how we proceed. This may include finding empirically supported solutions, e.g., forming minimal groupings that are more salient than automatic groupings we would like to avoid, like race. Ultimately, we should involve this scientific knowledge in our moral deliberations so that we can assess our moral choices with transparency and clarity. In doing so, we will increase our moral agency, making decisions that combine our moral values with a realistic picture of how we actually think.

CHAPTER 5: The Philosophical Value of an Informed Introspection

5.0 Moral Agency

By now, I hope to have persuaded the reader that scientific information can illuminate our moral cognitive processes and that there is value in learning this information because it can help us correct our moral behaviors to better align with our deeply held moral principles. This informed introspection gives us the opportunity to govern our moral behaviors in ways unavailable to us before. In situations where we may have been unknowingly motivated by instinct, desire and emotion, we now have the tools to govern our own lives by principles and values – this is the view of moral agency Korsgaard advances. This is the essence of morality; it is the enhancement of our human capacity for normative self-government.

Recall that Korsgaard developed her view from Kant's notion of autonomy: we are autonomous when we make our own laws. Korsgaard's autonomous agent legislates for herself what laws she wishes to uphold. We have intentions, we assess them, and we choose to adopt them. Our morality, then, is not a function of the content of our intentions, but rather a function of the exercise of our self-government (Korsgaard, 2006, p. 112). This is precisely what the informed introspection empowers us to do; it gives us the tools to effectively self-legislate. We uncover once-hidden cognitive processes, learning how they at times cause us to veer from the laws by which we have chosen to be governed. By illuminating the innerworkings of our moral behaviors, we can begin to evaluate them. We can hold them up to our moral principles and determine whether they inhibit or enable us to follow the moral laws we have chosen. And then we decide for ourselves whether we want to correct for them.

The autonomy that we are granted by the informed introspection contributes to the very essence of moral agency, and it is through this autonomy that our morality emerges. Morality is about our ability to form and act on judgments of what we ought to do. If we do not incorporate scientific information about our moral cognitive processes into our moral deliberations, our deliberative process will be underpowered, and we will fail to employ our full moral agency. If we are to adopt Korsgaard's views about the critical importance of normative self-government to our morality, we must not ignore the tools that can help us govern our own moral lives.

To conclude this dissertation, I bring two more philosophers into our discussion: John Dewey and John Rawls. My goal is to broaden the philosophical context for my view by engaging with their moral theories. Here, I examine two ethical views that I argue are compatible with and enlightened through my picture: Dewey's views of moral reflection and Rawls's notion of reflective equilibrium. Through this examination, I show that there is a harmony between my position and these well-established ethical views. Thus, I argue that we have good reason to believe that there is philosophical value in using scientific information about our moral cognition in our moral deliberations.

5.1 Dewey's *Moral Reflection*

> In principle a revolution was wrought when Hebrew prophets and Greek seers asserted that conduct is not truly conduct unless it springs from the heart, from personal desires and affections, or from personal insight and rational choice.
>
> – John Dewey
> *Ethics* 1936

In this section, I introduce Dewey's thoughts on moral reflection as well as his views about the role that science should play in our study of morality. Then I integrate Dewey's ideas with my own position, showing a strong harmony between the two.

In a revised edition of Ethics written with James H. Tufts, John Dewey discusses the nature of moral theory, highlighting the role of reflective morality. Moral theory, according to Dewey, can do three things: (i) generalize types of moral conflicts so that individuals can better clarify their own deliberations by seeing their problems in a larger context, (ii) state the leading ways that people (i.e., intellectuals) have dealt with various problems, and (iii) enhance personal moral reflection by suggesting alternatives that may otherwise be overlooked, which will lead to greater consistency in judgment (Dewey & Tufts, 1936, p. 175). Importantly, moral theory does not have ready-made conclusions for us; it does not command our choices. This would contradict the very nature of moral theory. What this means is that Dewey's treatment of moral theory and moral reflection is built on a process – moral theory is active. Dewey values the practical consequences of reflecting on our moral behavior: by understanding how we and others act in moral situations, we can see our moral process as a whole. This informs our moral conduct, allowing us to behave in ways that align with our moral values.

Dewey holds that no theory, moral or physical, can operate in a vacuum. He writes, "Moral as well as physical theory requires a body of dependable data, and a set of intelligible working hypotheses" (Dewey & Tufts, 1936, p. 190). Dewey thinks of four different areas where we can look to get dependable data for moral theory. I will briefly share those here, roughly out of order so that I can discuss the third area last, for reasons which will soon become clear.

First, we may get dependable data for our moral theory by looking at moral codes throughout tradition. These moral codes inevitably conflict, as they attempt to represent multiple

perspectives and cultures, changing as communities change. A dogmatist will pick out one of the many moral codes – probably the one that agrees most with his education and taste (Dewey & Tufts, 1936 p. 191). A reflective person will look at all these different codes as possible data, taking into account how they arose and were formed, as well as how they are relevant in the current day. The reflective position allows for these conflicting moral codes to be stored, not thrown away, so that we may pick from them what is important for our current ideas about what is good.

Second, we should consider things such as legal history, judicial decisions and legislative activity. This is closely connected to the first category about moral codes, since our moral codes are on display in these settings. The very purpose of courts and legislatures is to direct human conduct. By knowing the course of their deliberations, we can see how the moral codes held by justices and legislators influenced the arguments they stood by. Furthermore, we can inquire into the consequences of their judgments and laws. Dewey notes that we should also include more informal contributions here, such as biographies of those who have contributed moral teachings to the society.

Third, but really Dewey's fourth area, is theoretical models and conclusions from European and Asian history. Intellectuals from these great traditions have engaged in analysis and have developed directive principles based in rationality. He writes that at first, the vast number of different, logically incompatible positions may "indicate simply a scene of confusion and conflict" (Dewey & Tufts, 1936, p. 192). However, a closer study reveals that moral situations are greatly complex and no theory will be comprehensive enough – they will all leave out important factors. However, the goal here is not to look through history to find the theory or model that we wish to live by, or to combine the right parts of various theories to make some

eclectic combination. Instead, the proper inference to draw here is that "each great system of moral thought brings to light some point of view from which the facts of our own situations are to be looked at and studied" (Dewey & Tufts, 1936, p. 193). That is, these theories give to us questions that we can consider and apply to our current circumstances. These theories model for us what we should consider in our contemporary moral theories.

The remaining category of areas from which we can gain dependable data for our moral theory, is "found in the various sciences, especially those closest to man, such as biology, physiology, hygiene and medicine, psychology and psychiatry, as well as statistics, sociology, economics, and politics" (Dewey & Tufts, 1936, pp. 191-192). I will first let Dewey speak here, and then draw the clear connections to my view.

Dewey recognizes that the latter disciplines, which are more social in nature, differ from the former disciplines, insofar as they usually present more problems than solutions. However, even so, Dewey writes that it's a good thing for moral theory to get problems "more clearly in mind" and furthermore, "the very fact that these social disciplines usually approach their material independently of consideration of moral values has a certain intellectual advantage for the moralist" (Dewey & Tufts, 1936, p. 192). So these social sciences are important to our moral theory because there is some impartiality in their method – they approach their subject-matter in a way that is dethatched from moral convictions because they are aware that this may allow prejudices to seep in. Importantly, the biological and behavioral sciences give us "highly valuable techniques for study of human and social problems and the opening of new vistas" (Dewey & Tufts, 1936, p. 192). Here, Dewey gives an example of how scientific findings about public health have opened up a new body of moral interests and responsibilities. In the early 20th century, there was new scientific understanding of the consequences of poor hygiene and

sanitation, as well as how lack of nutrition is associated with disease, or how unsafe work environments can cause injury, and so on. Dewey's point here is that these scientific advancements created a new recognition of moral concerns. A parallel for today would be new moral concerns that emerge with our understanding of artificial intelligence, or with our astronomical discoveries. When science opens new doors, new matters will require our moral consideration.

I am in agreement with all of Dewey's views that I have presented here, and especially his thoughts about the role of scientific information in our moral theories. I believe that my view is compatible with Dewey's views. I believe that science can teach us about processes we would be otherwise unaware of, but these processes matter for our moral considerations. Just as scientific inquiry led us to understand hygiene better, allowing us to make better decisions about the sanitary habits we keep, scientific inquiry into our moral cognition will allow us to make better decisions about other types of habits we keep. We will find that new doors open with this knowledge, bringing new matters that will require our moral consideration. The cases I present in the following chapters will give examples of how scientific inquiry has illuminated our moral cognitive processes in ways that should change how we deliberate certain moral circumstances. Earlier in this chapter, I gave the example of in-group/out-group thinking. Knowing about my tendency to favor in-group members even if I know they might act badly, I should be careful in my deliberations involving in-group members. I should use this information that science has presented, as dependable data for my moral theories. These scientific findings cannot tell me which moral decisions to make, but they can help me reflect on moral situations in ways Dewey calls for. Greene's Trolley Problem experiments help us generalize a type of moral conflicts so that we can see them in a larger context. The in-group/out-group studies reveal that our moral

cognitive processes can be affected by factors we find to be morally irrelevant, suggesting alternative methods to correct for these biases.

It is crucial that we support our theories with data, which is precisely what a scientific understanding of our moral cognition allows us to do. In other writings, Dewey has further developed his views about science and morality, making a case specifically for the scientific investigation of morality.

In a pair of essays for The Philosophical Review in 1902, Dewey took up the question of how morality may be brought into the domain of science and what bearing a scientific treatment may have on morality itself. He finds that only a historical method, also referred to as an evolutionary or genetic method, would provide a sufficient scientific examination of morality. This historical approach would "not only enable us to interpret both its [morality's] cruder and more mature forms, but – what is even more important – would give us insight into the operations and conditions which make for morality, and thus afford us intellectual tools for attacking moral facts" (Dewey, 1902a, p. 124). This would have practical consequences for our moral thinking. Scientific investigations in other domains, e.g. physics, give us results that we may implement in order to get desirable consequences. Likewise, Dewey explains, knowledge about the generation of morality will give us insight into the innerworkings of morality, which can provide us a path for practical control.

Furthermore, Dewey finds that this evolutionary method "unites the present situation with its accepted customs, beliefs, moral ideals, hopes, and aspirations, with the past. It sees the moral process as a whole, and yet in perspective" (Dewey, 1902b, p. 370). In this same spirit, I argue that our scientific knowledge of our moral cognition and how it has evolved will inform our current moral conduct, making it significant for our moral deliberations. Furthermore, this

knowledge will allow us to make moral decisions with which we are reflectively satisfied – not because we're making the "right" moral decisions, but because we are better able to make moral decisions that we are more satisfied with in retrospect.

5.2 Challenging Rawls's *Reflective Equilibrium*

An important notion within John Rawls's famed justice as fairness theory is reflective equilibrium. He introduces this concept as follows:

> According to the provisional aim of moral philosophy, one might say that justice as fairness is the hypothesis that the principles which would be chosen in the original position are identical with those that match our considered judgments and so these principles describe our sense of justice. But this interpretation is clearly oversimplified… From the standpoint of moral philosophy, the best account of a person's sense of justice is not the one which fits his judgments prior to his examining any conception of justice, but rather the one which matches his judgments in reflective equilibrium (Rawls, 1971, p. 48).

Rawls rejects the notion that our moral philosophy should rest on foundational beliefs; instead, he looks for a theory of justice that is based upon principles of fairness that we, upon reflection, believe should govern us.[27] Self-examination is important to Rawls's proposal. He considers moral philosophy to be Socratic, pointing out that once we see our moral principles brought to light, i.e., exhibited in particular cases about which we form moral judgments, we may want to reconsider them. Similarly, when we understand our moral principles better, we may reconsider the moral judgments we make. So, Rawls argues that we should engage in the method of reflective equilibrium to properly examine our beliefs of all kinds, moral and non-moral alike. Below, I describe Rawls's method, which I believe is compatible with my own view. To show its

[27] Some philosophers believe Rawls's account to be compatible with foundationalist theories. See Daniels (2020) for further analysis.

compatibility, I offer a friendly amendment to this method, showing that it will improve our process of reflective equilibrium.

In our reflection, we consider all kinds of beliefs and adjust them with one another until we find an equilibrium, where they are coherent with one another. These beliefs, in addition to being of various kinds (e.g., political, moral, inductive, scientific) also range from general to specific. Rawls suggests two methods to accommodate the different ways in which we reach coherence among our beliefs: narrow reflective equilibrium and wide reflective equilibrium.

In narrow reflective equilibrium, we consider our specific judgments and how they fit with our principles. For example, a utilitarian engaging in narrow reflective equilibrium considers her judgments about the Trolley Problem, where she may feel a tussle between her utilitarian principles and her moral judgments about how to respond in different variations of the moral dilemma. In the switch case, she feels no imbalance between her judgment that flipping a switch to save five lives at the cost of one, is the right decision. In the footbridge case, however, our utilitarian may hesitate at the idea of pushing someone over a bridge to save the lives of others, and she feels a little more resistance between her judgments and principles. Other variations of the Trolley Problem ask us to consider our family members as the ones we are meant to sacrifice and strangers as those who we would save – we can imagine this causes our utilitarian more imbalance between her judgments and her principles. So, in engaging in reflective equilibrium, she adjusts both her judgments to these specific versions of the Trolley Problem, and she adjusts her utilitarian principles. Perhaps she adjusts her greatest-happiness-for-greatest-number principle to account for the additional pain that would be caused by a person ending the life of their own kin. Perhaps she adjusts her never-throwing-someone-off-a-footbridge judgment by promising herself that if she were ever in that position, she would try to

override her emotional response and do what needed to be done to save the innocent people. In engaging in this process, our utilitarian is finding cohesion between her judgments of specific questions with her moral principles.

In wide reflective equilibrium, we consider our principles and theories with alternative principles and theories. It is unrealistic to expect that we would consider all alternatives, since there would be no conceivable end to that process. Rawls writes, "The most we can do is to study the conceptions of justice known to us through the tradition of moral philosophy and any further ones that occur to us, and then to consider these" (Rawls, 1971, p. 49). Thus, we should ask general questions, comparing our beliefs with those that belong to other theories. Our utilitarian would engage in wide reflective equilibrium if she deliberates whether she should keep her utilitarian beliefs or adopt deontological beliefs. She may step back from the Trolley Problem variations for just a moment, asking herself whether it is right to value the greatest-happiness-for-the-greatest-number or whether some a deontological belief, e.g., never-treat-others-as-means-to-an-end should be adopted. In her wide reflective process, she may consider her judgments about the Trolley Problem variations, building cohesion between her judgments and the wider variety of principles she adopts. Importantly, though, in her reflection she will ask herself how these two theories compare with one another and which principles she should adopt.

Whether in the narrow or wide reflective equilibrium process, it is important that the moral agent goes back and forth between her beliefs, judgments and tentative principles, correcting them to find coherence. Here, at the end of her deliberative process, our moral agent will find equilibrium. Crucially, though, this equilibrium is not stable and may require further deliberation in the future. When an imbalance arises, Rawls points out that "the important thing is to find out how often and how far [our theory] is wrong. All theories are presumably mistaken

in places. The real question at any given time is which of the views already proposed is the best approximation overall" (Rawls, 1971, p. 52). So, in building and maintaining our moral philosophy, we should understand our beliefs to be revisable at any time.

I am in agreement with Rawls's reflective equilibrium. I too reject the notion that our moral principles should rest on a select few foundational beliefs that are immune to revision. My view does not seek to change Rawls's reflective process, only to interrupt it before it begins. For Rawls, the method of reflective equilibrium is an interpretive process which allows us to reject various kinds of beliefs if we are unable to maintain coherence between them and our other beliefs. So, the utilitarian rejects her never-throwing-someone-off-a-footbridge judgment because it doesn't fit with her utilitarian principles. This is a rejection based on convenience, and one that she need not justify further.[28]

Here is my suggestion: before we engage in the Rawlsian method of reflective equilibrium, we should reflect on our beliefs insofar as they are coherent with our scientific knowledge. Our understanding of moral cognition informs us that sometimes our moral judgments are triggered by factors that we ourselves deem to be morally irrelevant. In a scientifically informed moral reflection, we reconcile our moral beliefs and actions with our moral principles. This accounting should be completed before we get to a Rawlsian reflection. After we understand our beliefs and reject those that we have justified reasons against (i.e., they are incompatible with our moral principles), then we can engage with the method of reflective equilibrium. At this point, the beliefs still in our consideration will be ones that have passed the

[28] Rawls's theory is taken by some readers as a descriptive project and others as a justificatory project, and for the purposes of this evaluation, it does not matter which interpretation we choose.

scientific assessment. Thus, when we are balancing our beliefs in hopes of finding equilibrium, we are doing so not based on convenience alone but also on scientific grounding.

5.3 Conclusion

In this dissertation, I have sought to lay a foundation for my view of the role science should play in our moral deliberations. In Chapter 1, I explored four ways in which Kitcher proposed sociobiologists attempt to "biologicize" ethics. What I found here was that the naturalistic fallacy is a widespread and tempting problem for scientists interested in ethical questions. Although there were some promising notes in Kitcher's taxonomy, I do not feel that his picture captured what I believe to be an important role that science can play in morality. So, I proposed an addition to his list. My addition holds that sociobiology can teach us about how our moral intuitions may drift away from our moral commitments, and this inquiry allows us to make practical changes to increase coherence between our moral thinking and our moral action, increasing satisfaction in retrospect. Importantly, my view does not commit the naturalistic fallacy, as it is not interested in making normative claims based on scientific findings. Instead, science contributes facts about our moral cognition that would otherwise be hidden to us. My "ought" lies in my appeal to the practical consequences our scientific findings can have on our moral deliberations. It is a philosophical appeal that claims that implementing this knowledge will bring the moral agent overall satisfaction in the long-term.

After making my proposal, I considered the compatibility of my theory with Korsgaard's view on moral agency. Korsgaard views morality as a human practice grounded in reason. Her view of normative self-government holds that humans, as autonomous agents, choose the laws

that govern us, and this is from where morality emerges. I believe that building our knowledge of how our moral cognition works contributes to our autonomy, allowing us to better practice normative self-government.

In Chapter 2, I examined two naturalistic accounts of morality that I think go too far by attempting to use science to uncover what we should value. Both Kitcher and Greene inflate the role of science in moral theory by drawing normativity from their descriptive theories. This, I argued, violated the is/ought boundary. Even though neither theory directly invokes supernatural beings, they each employ a natural proxy for an external moral authority. In doing so, Kitcher and Greene commit the naturalistic fallacy.

Then, in Chapter 3, I introduced the idea of an informed introspection. I addressed the empirical question of whether including scientific information in our moral deliberations would be effective. I considered two ways in which it could backfire, but ultimately found reason for optimism from the empirical research itself. Then, I examined two concrete cases in which our scientific investigations reveal ways in which our moral intuitions come apart from our moral commitments. In both these cases, I showed that factors that we ourselves may consider to be morally irrelevant are influencing our moral judgments and actions. I argued that we include this scientific in our moral deliberations not because the science tells us what answer we should make but because the science informs us in a way that empowers us to make informed decisions about our moral behaviors. This does not only give us more information and power; it will give us more satisfaction. By learning to make moral decisions that better align with our moral principles, we will be able to better self-govern and we will be more satisfied in our moral reflections.

In Chapter 4, I continued my proposal of the informed introspection, showing that we should use it to be wary of our moral intuitions. Because our unconscious, automatic processes are very quick to categorize our social environment, we should be aware of the deep influences this has on our attitudes and preferences of others. If we value fairness, for example, it is vital to understand how simple – morally irrelevant – groupings, like the color of a sticker one wears, can result in biases that favor our ingroup and disfavor our outgroup. Then, if we find that this is unacceptable because it is incoherent with our moral principle, we can choose to correct for it in our moral deliberations. I reviewed practical ideas from the empirical studies that may help us do just this: for example, re-forming the group membership to reflect ingroup/outgroup distinctions that we find acceptable. At the very least, we can incorporate our knowledge of this tendency to build skepticism of our moral intuitions into our deliberative process. Being cautious of our moral intuitions can help us adjust our decision-making in a way that is coherent with our deeply held moral principles.

In the present chapter, I revisited Korsgaard's notion of normative self-government, joining it with the informed introspection. I showed that the informed introspection is central to our moral agency, and thus to our very morality. I also brought two additional philosophers into the discussion. I found that Dewey's view of moral reflection was very harmonious with my own view. In fact, in Ethics, Dewey describes four different areas that can give us dependable data that we can use in our moral theory. I argued that my view fits within one of these ways. Furthermore, Dewey has also defended the scientific investigation of morality, which is compatible with my theory. From Dewey, we get both that (i) he believed science can give us data that is important to our moral theories and that (ii) as a method, science can help us better understand morality. Both of these claims support the view I propose.

Furthermore, Rawls's method of reflective equilibrium is improved by my view. Although I agree with much of what Rawls's method offers, I have reservations about the basis upon which we are able to discard our beliefs. While trying to build coherence between our judgments, principles and theories, we can reject certain judgments if they are inconvenient to us. Ideally, we will do so carefully, but nonetheless we are motivated by our desire to achieve and maintain an equilibrium. My view offers an initial filter that will catch some corrupted beliefs before we can entertain them in Rawls's reflective method. By understanding our moral cognition, we will be able to identify ways in which our moral intuitions lead us astray. We can use this scientific information in our moral deliberations, catching intuitions that we believe to be confounded by morally-irrelevant cognitive processes and biases. If we deem our intuitions and judgments to be well-vetted, then we can pass them along to the method of reflective equilibrium, where we can further examine their coherence with our moral principles and theories. This is an improvement to Rawls's picture because our reflections are no longer based solely on convenience, but now also on our scientific knowledge.

In conclusion, I believe there is practical value in using scientific information about our moral cognition in our moral deliberations. In doing so, we increase our autonomy over our own moral behaviors, making us better moral agents and allowing us to live up to the morals we value.

REFERENCES

Allport, G. W. (1979). *The Nature of Prejudice*. Perseus Books.

Apfelbaum, E. P., Sommers, S. R., & Norton, M. I. (2008). Seeing race and seeming racist? Evaluating strategic colorblindness in social interaction. *Journal of Personality and Social Psychology, 95*(4), 918–932.

Apfelbaum, E. P., Pauker, K., Sommers, S. R., & Ambady. N. (2010). In Blind Pursuit of Racial Equality? *Psychological Science, 21*(11), 1587–1592. https://doi.org/10.1177/0956797610384741

Apfelbaum, E. P., Norton, M. I., & Sommers, S. R. (2012). Racial Colorblindness: Emergence, Practice, and Implications. *Current Directions in Psychological Science 21*(3), 205–209.

Ashburn-Nardo, L., Knowles, M. L., & Monteith, M. J. (2003). Black Americans' implicit racial associations and their implications for intergroup judgment. *Social Cognition 21,* 61–87.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). The Guilford Press.

Baron, A. S. (2015). Constraints on the Development of Implicit Intergroup Attitudes. *Child Development Perspectives, 9,* 50-54. https://doi.org/10.1111/cdep.12105

Baron, A. S., & Banaji, M. R. (2006). The Development of Implicit Attitudes: Evidence of Race Evaluations From Ages 6 and 10 and Adulthood. *Psychological Science, 17*(1), 53–58. https://doi.org/10.1111/j.1467-9280.2005.01664.x

Baron, A. S., & Dunham, Y. (2015). Representing 'Us' and 'Them': Building Blocks of Intergroup Cognition, *Journal of Cognition and Development, 16*(5), 780-801. DOI: 10.1080/15248372.2014.1000459

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323–370.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *A dual process model of impression formation* (pp. 1–36). Lawrence Erlbaum Associates, Inc.

Brownstein, M. (2019). Implicit Bias. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/.

Cantor, J. M., Kabani, N., Christensen, B. K., Zipursky, R. B., Barbaree, H. E., Dickey, R., Klassen, P. E., Mikulis, D. J., Kuban, M. E., Blak, T., Richards, B. A., Hanratty, M. K.,

& Blanchard, R. B. (2008). Cerebral white matter deficiencies in pedophilic men, *Journal of Psychiatric Research, 42*(3), 167-183. https://doi.org/10.1016/j.jpsychires.2007.10.013.

Chernyak, N., Harris, P. L., & Cordes, S. (2018). Explaining early moral hypocrisy: Numerical Cognition promotes equal sharing behavior in preschool-aged children. *Developmental Science, 22*(1). https://doi.org/10.1111/desc.12695

Ciaramelli, E., Muccioli, M., Làdavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 2*(2), 84–92. https://doi.org/10.1093/scan/nsm001

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. G.P. Putnam.

Daniels, N., (2020). Reflective Equilibrium. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/.

Dewey, J. (1902a). The evolutionary method as applied to morality. *The Philosophical Review, 11*(2), 107. doi:10.2307/2176631

Dewey, J. (1902b). The evolutionary method as applied to morality: ii. its significance for conduct. *The Philosophical Review, 11*(4), 353. doi:10.2307/2176470

Dewey, J., & Tufts, J. H. (1936). *Ethics* (Revised ed.). Holt.

Dunham, Y., Chen, E., & Banaji, M. R. (2013). Two signatures of implicit intergroup attitudes: Developmental invariance and early enculturation. *Psychological Science, 24*(6), 860–868.

Dunham, Y., Newheiser, A., Hoosain, L., Merrill, A., & Olson, K. R. (2014). From a different vantage: Intergroup attitudes among children from low- and intermediate-status racial groups. *Social Cognition, 32*, 1–21.

Fiske, S.T., & Neuberg, S.L. (1990). A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. *Advances in Experimental Social Psychology, 23*, 1-74.

Foot, P. (1967). The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review, 5*, 5-15.

Goldstick, J. E., Cunningham, R.M., & Carter, P.M. (2022). Current causes of death in children and adolescents in the United States. *New England Journal of Medicine, 386*, 1955-1956. DOI: 10.1056/NEJMc2201761

Goodwin, G. P., & Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition, 106*(3), 1339–1366. https://doi.org/10.1016/j.cognition.2007.06.007

Greene, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science, 293*, 2105-2108.

Greene, J. D., Nystrom, L.E., Engell, A.D., Darley, J.M. & Cohen, J.D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389-400.

Greene, J. D., Morelli, S.A., Lowenberg, K., Nystrom, L.E. & Cohen, J.D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144-1154.

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology, 45*(3), 581–584. https://doi.org/10.1016/j.jesp.2009.01.003

Greene, J. D. (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin Books.

Greene, J. D. (2014). Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. *Ethics, 124*(4), 695-726.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review 108*(4), 814–34. https://doi.org/10.1037/0033-295x.108.4.814.

Haidt, J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Vintage Books.

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me = bad: infants prefer those who harm dissimilar others. *Psychological Science, 24*(4), 589–594. https://doi.org/10.1177/0956797612457785

Hewstone, M., Hantzi, A., & Johnston, L. (1991). Social categorization and person memory: The pervasiveness of race as an organizing principle. *European Journal of Social Psychology, 21*, 517-528.

Hirschfeld, L. A. (1996). *Race in the Making: Cognition, culture, and the child's construction of human kinds.* MIT Press.

Hirschfeld, L. A. (2001). On a folk theory of society: Children, evolution, and mental representations of social groups. *Personality and Social Psychology Review, 5*(2), 107–117.

Holmes, M. (2023, January 9). What is child endangerment? When leaving your child alone becomes a crime. *Huffington Post*. Retrieved May 18, 2023, from https://www.huffpost.com/entry/what-is-child-endangerment_l_63b73bdae4b0b2e1506638e8.

Hume, D. (1777). *An enquiry concerning the principles of morals.* Project Gutenberg, January 12, 2010. https://www.gutenberg.org/files/4320/4320-h/4320-h.htm

Hume, D. (2000). *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects* (Norton, M. J., & Norton, D. F., Eds.). Oxford University Press.

Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology, 85*, 616-626.

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore, *Research in Organizational Behavior, 29,* 39-69. https://doi.org/10.1016/j.riob.2009.10.001.

Kitcher, P. (2011). *The Ethical Project.* Harvard Univ Press.

Kitcher, P. (2006). Four Ways of Biologicizing Ethics. In E. Sober (Ed.), *Conceptual Issues in Evolutionary Biology* (pp. 575-586). MIT Press.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908–911. https://doi.org/10.1038/nature05631

Korsgaard, C. (2006). Morality and the Distinctiveness of Human Action. In F. de Waal (Ed.), *Primates and Philosophers* (pp. 98-119). Princeton University Press.

Korsgaard, C. (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press.

Latané, B., & Rodin, J. (1969). A lady in distress: Inhibiting effects of friends and strangers on bystander intervention. *Journal of Experimental Social Psychology, 5*(2), 189–202. https://doi.org/10.1016/0022-1031(69)90046-8

Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). The Origins of Social Categorization. *Trends Cognitive Science, 21*(7), 556-568. doi: 10.1016/j.tics.2017.04.004.

Liptak, A. (2022, April 25). Supreme court leans toward coach in case on school prayer. *New York Times*. Retrieved May 18, 2023, from https://www.nytimes.com/2022/04/25/us/politics/supreme-court-prayer-football-coach.html.

Liu, B., & Ditto, P. H. (2012). What Dilemma? Moral Evaluation Shapes Factual Belief. *Social Psychological and Personality Science, 4*(3). https://doi.org/10.2139/ssrn.2071478

Luttrell, M., & Robinson, P. (2007). *Lone Survivor: The Eyewitness Account of Operation Redwing and The Lost Heroes of SEAL Team 10*. Back Bay Books.

Margaritoff, M. (2023, January 4). ABC news' dax tejera's wife faces charges for leaving kids unattended on night he died. *Huffington Post*. Retrieved May 18, 2023, from https://www.huffpost.com/entry/abc-news-dax-tejera-wife-criminally-charged_n_63b59217e4b0fe267cac8186 .

Marsh, A. A., Stoycos, S. A., Brethel-Haurwitz, K. M., Robinson, P., VanMeter, J. W., & Cardinale, E. M. (2014). Neural and cognitive characteristics of extraordinary altruists. *Proceedings of the National Academy of Sciences, 111*(42), 15036–15041. https://doi.org/10.1073/pnas.1408440111

McConnell, A. R., and Leibold, J. M. (2001). Relations among the Implicit Association Test, explicit attitudes, and discriminatory behavior. *Journal of Experimental Social Psychology 37*, 435–442.

McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology, 45*(3), 577–580. https://doi.org/10.1016/j.jesp.2009.01.002

Mendez, M. F., Anderson, E. K., & Shapira, J. S. (2005). An Investigation of Moral Judgement in Frontotemporal Dementia. *Cognitive and Behavioral Neurology 18*, 193-197.

Mikkelsen, K., Stojanovska, L., Polenakovic, M., Bosevski, M., & Apostolopoulos, V. (2017). Exercise and mental health. *Maturitas, 106*, 48–56. https://doi.org/10.1016/j.maturitas.2017.09.003

Moore, A., Clark, B., & Kane, M. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. Psychological Science, 19(6), 549–557.

Moore, A., Lee, N., Clark, B., & Conway, A. (2011). In defense of the personal/impersonal distinction in moral psychology research: Cross-cultural validation of the dual process model of moral judgment. *Judgment and Decision Making, 6*(3), 186-195. doi:10.1017/S193029750000139X

Newheiser, A. K., & Olson, K. R. (2012). White and Black American children's implicit intergroup bias. *Journal of Experimental Social Psychology, 48,* 264–270.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101

O'Connor, C., Fulton, N., Wagner, E. & Stanford, P. K. (2012). Deus Ex Machina: A Cautionary Tale for Naturalists. *Analyse & Kritik 34*(1). doi:10.1515/auk-2012-0104.

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology, 42,* 1051-1068.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in Cognitive Sciences, 11,* 37–43.

Pronin, E. (2009). The Introspection Illusion. *Advances in Experimental Social Psychology, 41*.

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology, 43*, 565–578.

Rawls, J. (1971) *Theory of Justice.* The Belknap Press of Harvard University Press.

Rawls, J. (2001) *Justice as Fairness: A Restatement*. The Belknap Press of Harvard University Press.

Roberts, S. O., Gelman, S. A., & Ho, A. K. (2017). So it is, so it shall be: Group regularities license children's prescriptive judgments. *Cognitive Science. 41*(3), 576-600. Doi: 10.1111/cogs.12443.

Roberts, S. O., Guo, C., Ho, A. K., & Gelman, S. A. (2018). Children's descriptive-to-prescriptive tendency replicates (and varies) cross-culturally: Evidence from China. *Journal of Experimental Child Psychology 165,* 148-160.

Roberts, S. O., Ho, A. K., & Gelman, S. A. (2019). The Role of Group Norms in Evaluating Uncommon and Negative Behaviors. *Journal of Experimental Psychology 148*(2), 374-387.

Roberts, S. O., & Horii, R .I. (2019). Thinking Fast and Slow: Children's Descriptive-To-Prescriptive Tendency Under Varying Time Constraints. *Journal of Cognition and Development 20*(5), 790-799.

Roberts, S.O., Ho, A. K. & Gelman, S. A. (2020). Should Individuals Think Like Their Group? A Descriptive-to-Prescriptive Tendency Toward Group-Based Beliefs. *Child Development, 92.* doi: 10.1111/cdev.13448

Rudman, L. A. (2004). Social justice in our minds, homes, and society: The nature, causes, and consequences of implicit bias. *Social Justice Research, 17*, 129-142.

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues 57,* 743–762.

Rudman, L. A., & Lee, M. R. (2002). Implicit and explicit consequences of exposure to violent and misogynous rap music. *Group Processes & Intergroup Relations*, *5,* 133–150.

Ruse, M., & Wilson, E.O. (1986). Moral Philosophy as Applied Science. *Philosophy, 61,* 173–192.

Sandel, M. J. (2009). *Justice: What's the Right Thing to Do*? Farrar, Straus and Giroux.

Sapolsky, R. M. (2017). *Behave: The Biology of Humans At Our Best and Worst*. Penguin Books.

Simon, J. C., & Gutsell, J. N. (2020). Effects of minimal grouping on implicit prejudice, infrahumanization, and neural processing despite orthogonal social categorizations. *Group Processes & Intergroup Relations*, *23*(3), 323–343. https://doi.org/10.1177/1368430219837348

Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010). Moral intuitions. In J. M. Doris (Ed.) & Moral Psychology Research Group, *The moral psychology handbook* (pp. 246–272). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199582143.003.0008

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*(6), 895–917. https://doi.org/10.1037/0022-3514.88.6.895

Stangor, C., Lynch, L., Duan, C., & Glass, B. (1992). Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology, 62*, 207-281.

Tajfel, H., Billig, M., Bundy, R., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology, 1*, 149-178.

Tavernise, S. (Host). (2022, April 27). The Supreme Court Considers a Football Coach's Prayers. [Audio podcast episode]. *The Daily. New York Times.* https://www.nytimes.com/2022/04/27/podcasts/the-daily/school-prayer-supreme-court.html

Thomas, A. J., Stanford, P. K., & Sarnecka, B. W. (2016). No Child Left Alone: Moral Judgments about Parents Affect Estimates of Risk to Children. *Collabra, 2*(1). https://doi.org/10.1525/collabra.33

Thomson, J. (1985) The Trolley Problem. *The Yale Law Journal 94*(6), 1395-1415.

Turner, J. C., Brown, R. J., & Tajfel, H. (1979). Social comparison and group interest in ingroup favouritism. *European Journal of Social Psychology, 9*(2), 187–204. https://doi.org/10.1002/ejsp.2420090207

Valdesolo, P., & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science, 17*(6): 476-477.

Van Bavel, J. J., & Cunningham, W. A. (2009). Self-Categorization With a Novel Mixed-Race Group Moderates Automatic Social and Racial Biases. *Personality and Social Psychology Bulletin, 35*(3), 321–335. https://doi.org/10.1177/0146167208327743

Van Bavel, J. J., FeldmanHall, O., & Mende-Siedlecki, P. (2015). The neuroscience of moral cognition: From dual processes to dynamic systems. *Current Opinion in Psychology, 6*, 167-172.

Wilson, E. O. (1975). *Sociobiology: The New Synthesis*. Harvard University Press.

Wilson, E. O. (1978). *On Human Nature*. Harvard University Press.

Wright, J. C., Cullum, J., & Schwab, N. (2008). The cognitive and affective dimensions of moral conviction: Implications for attitudinal and behavioral measures of interpersonal tolerance. *Personality and Social Psychology Bulletin, 34*(11), 1461–1476. https://doi.org/10.1177/0146167208322557