2019

The Problems Of Moral Psychology

Thomas Noah
University of Pennsylvania

The Problems Of Moral Psychology

Abstract

This dissertation is a collection of three essays centered around outstanding fundamental problems in the field of moral psychology. These fundamental problems concern both metaphysical and methodological disagreements – namely, what is the subject-matter of moral psychology? And what are the methods for the investigation of that subject-matter? The first chapter examines the problem of marking the domain of moral psychology by isolating its subject-matter and the various methodologies for investigating it. By building from a minimal core of shared agreement, researchers should be able to classify different branches of moral psychology by both subject-matter and method of investigation while being quietist about the correct methodology. This is turn allows for the construction of a taxonomy of moral psychological approaches that allows researchers to efficiently locate the direct source of disagreements. The second chapter examines Lawrence Kohlberg's research program and identifies a particular assumption that guides that program while blocking further progress in the field. That assumption concerns the relation between normative theorizing and descriptive categorization, such that the explanations for why one moral theory is superior to another mirrors stages of moral development. By abandoning this assumption and other assumptions in its local vicinity, researchers could make progress without being bogged down in first-order normative disagreement. The third chapter looks to a recent debate concerning whether neuroscience is normatively significant. Against a standard interpretation of the debate in Anglophone philosophy that the argument for the normative insignificance of neuroscience is sound, I argue that the critique is only partly successful and not for the reasons commonly recognized. Rather than object to the program of demonstrating the normative significance of neuroscience on normative grounds, we ought to object to the program on descriptive grounds. Each chapter proceeds through arguments rooted in philosophical analysis and reflections on findings in the social and natural sciences – in particular, history, psychology, economics, sociology, anthropology, cognitive science, and neuroscience.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Philosophy

First Advisor

Cristina . Bicchieri

Keywords

Disagreement, Interdisciplinarity, Methodology, Moral Psychology, Normativity, Philosophy

Subject Categories

Philosophy

THE PROBLEMS OF MORAL PSYCHOLOGY

Thomas Noah

A DISSERTATION

in

Philosophy

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

_____

Errol Lord

Associate Professor of Philosophy

Graduate Group Chairperson

_____

Errol Lord, Associate Professor of Philosophy

Dissertation Committee

Cristina Bicchieri, S. J. Patterson Harvie Professor of Social Thought and Comparative Ethics

Adrienne Martin, Akshata Murty ’02 and Rishi Sunak Associate Professor of Philosophy, Politics and Economics and George R. Roberts Fellow

THE PROBLEMS OF MORAL PSYCHOLOGY

Dedication page

To Deloris, Glenda, Barbara, Glenda Jean, and Courtney

ACKNOWLEDGMENT

This collection of essays is the result of many years of co-deliberating with some of the finest philosophers and individuals I have been able to work with.

I thank my adviser Cristina Bicchieri for her continued support. Much of the work in the following essays comes out of work that I originally created for courses that she taught on empirical moral psychology, including a Penn-Rutgers-Princeton course co-taught with Stephen Stich and Gilbert Harman. She has been a constant source of inspiration for her attention to the intersections of empirical sciences with philosophical theories of behavior and human action, as well as her rigor and no-nonsense approach to philosophy. If not for her, I would not have received the training in neuroscience that made the third essay possible. And the opportunities she provided to me to work with UNICEF and other NGOs on social change proved important in broadening my understanding of the stakes of characterizing the domain of moral thought and behavior. We worked together over many years on my projects, hers, and joint projects. She was a very good adviser and is my friend.

I thank the members of my dissertation committee – Adrienne Martin and Errol Lord. They each have an acute sense of the moral stakes involved in the essays that follow, regardless of any differences I perceive between the philosophical traditions I understand them to work within. They are each remarkably good philosophers and mentors.

The essays thinking about the domain of moral psychology and the inheritance from Kohlberg were partly inspired by thinking through the implications of some of the lessons I learned from Adrienne in both a prior

metaethics class and through our personal interactions before she departed from the University of Pennsylvania. Even after leaving, she has always been generous and intellectually rigorous in our interactions and displayed a deep understanding of exactly the sorts of issues I attempt to unpack (even when my understanding was considerably less than).

Errol has been of immeasurable help to me in many ways. In particular, the third essay – concerning the relation between normative ethical theorizing, neuroscience, and psychology – developed the most through private meetings and extensive commentary from him. Errol was also instrumental in helping me to secure a spot at a National Endowment of the Humanities summer institute on moral psychology and moral education that helped me crystalize many of the issues of concern to me. Errol has been a model philosopher for his temperament and ability to see not only the arguments but also the rhetorical and dialectical spaces available to me. I regret that I did not work more closely with him throughout.

I thank Andrew McAninch, who served on my preliminary candidacy committee and who was quite valuable working out the dialectic and rhetorical strategy employed in the first chapter.

The original inspiration for this line of inquiry traces back to my time as an Ethical, Social, and Political Philosophy major at University of Massachusetts Boston. My undergraduate advisers Larry Blum and Lisa Rivera had a tremendous impact on my education, philosophical orientation, and work. I learned from each in their own ways how to attend to arguments and the spaces between arguments. But most importantly I learned that philosophy must attend

to real-world issues of normative significance. I hope that this collection of essays in some way honors their legacies, however imperfectly.

I continued to learn more about the nature of normative ethics while studying at Texas Tech University prior to coming to Penn. There, Danny Nathan, Howard Curzer, and Walt Schaller were especially vital to my development. I thank each of them for offering courses and counseling that allowed me to greater understand the varieties of ethical theories and their interactions with choice and behavior.

Many other philosophers have been important in my intellectual development, even if I did not work with them in a strictly mentoring relationship. From interactions and classes with them, though, I did learn a great deal about how to think through arguments and how to exemplify many epistemic virtues of what I consider to be good philosophy (although, admittedly, I often fail in this regard as well). At UMass Boston, Larry Kaye, Adam Beresford, and Yumiko Inukai were especially important to my intellectual development. At Texas Tech, Francesca di Poppa, Ed Averill and Jeremy Schwartz taught me a great deal. And at Penn, my final philosophical home, Gary Hatfield, Paul Guyer, Liz Camp, Susan Sauve Meyer, Michael Weisberg, Scott Weinstein, KC Tan, and Dan Singer have all served as various models for good philosophy in their own unique ways. In particular, I must thank Gary for running my first year proseminar on the history of analytic philosophy. I have partly absorbed his kind of "Drebenated" approach to philosophy, along with a certain kind of historical sense and reasonable skepticism.

Two other philosophers at Penn have been very important to me: Samuel Freeman and Lisa Mirrachi. I took some coursework with Samuel, and it (of course) was both on Rawls and very good – with a sensitive attending to arguments, history, good sense, and the variety of positions available in the conceptual space. And I was able to partly sit in on an early course from Lisa when she came to Penn. But their importance to me over time has been more personal. They helped me work through various difficulties that I experienced in graduate school and helped fulfill a role that the late Janet Farrell Smith played for me at UMass Boston. They are all great philosophers (exceptionally rigorous, philosophically no-nonsense, sensitive, and so on), but the special role was one of concern for me as an individual human being in an extra-academic way. I treasure their care and concern for helping me to work through various issues and difficulties that came about during my time and hope that others who need that kind of assistance can find it from those who have the ability and sensibility.

At Penn, I also received training and a graduate certificate in social, cognitive, and affective neuroscience. This was most helpful in working through the third essay. Martha Farah, who created and runs the program, has provided an invaluable service to students from various disciplines seeking to integrate the empirical study of the brain into their research, and I am deeply grateful for her patience and knowledge.

Many of my fellow students were also extremely helpful in working though my thought and research over the years. My closest co-deliberators (alternatively, my Philosophy Partners East and West) are Molly Sinderbrand and Kyle Adams, without whom much of this would not have been possible and

who share much of my philosophical sensibility but few of my vices. I also thank Charles Lloyd Phillips, with whom I discussed many earlier versions of some of the arguments contained in the collection. I also learned a great deal from members of my original cohort at Penn: Aditi Chaturvedi, Carlos Santana, and Justin Bernstein. Finally, I thank other fellow students who have helped my thinking over the years even if I failed to properly appreciate it at the time: Doug Paletta, Mike Nance, Paul Franco, Marcy Latta, Weibke Deimling, Chris Melenovsky, Alkistis Elliot-Graves, Rob Hoffman, Emily Park, Max Robitzsch, Lindsey Fiorelli, Hal Parker, Karen Kovaka, Collin Anthony, Rob Willison, Louise Daoust, Devin Curry, Nabeel Hamid, Jordan Taylor, Shereen Chang, Ben Baker, Patrick Ball, Marie Barnett, Chetan Cetty, Daniel Fryer, Raj Patel, Pierce Randall, and Brian Reese.

In my non-academic sphere, I thank my family – especially my mother, my sister, and my grandparents. They are very special to me in so many ways that to speak more would be a disservice.

Former administrative assistants Gerri Winters and Sandy Natson provided me so much warmth, friendship, and assistance during our overlapping time at Penn. Even though they are now in retirement, I cherish them. They made me feel part of a community, and I could always rely on them for a laugh, a cry, a sympathetic ear, and counsel that exemplified the best virtues of their faith traditions.

Finally, I thank my ex-wife for all the support that she provided over the years. We began a long journey together back in the early 2000s that took us around the country to various institutions as we both pursued our goals and

ground projects. Although no longer together and although we grew apart eventually, this project literally is the result of many hours of her sacrifice, love, and skepticism toward the pretenses of philosophers. I thank her greatly for everything.

For anyone else who I have failed to mention explicitly, I thank you for whatever contribution you provided to me. It was difficult to remember all those who were instrumental in getting me to this location, and I should have kept better track along the way. I deeply apologize.

In the end, though, this project and the faults with it are mine. I own the faults in their entirety.

ABSTRACT

THE PROBLEMS OF MORAL PSYCHOLOGY

Thomas Noah

Cristina Bicchieri

This dissertation is a collection of three essays centered around outstanding fundamental problems in the field of moral psychology. These fundamental problems concern both metaphysical and methodological disagreements – namely, what is the subject-matter of moral psychology? And what are the methods for the investigation of that subject-matter? The first chapter examines the problem of marking the domain of moral psychology by isolating its subject-matter and the various methodologies for investigating it. By building from a minimal core of shared agreement, researchers should be able to classify different branches of moral psychology by both subject-matter and method of investigation while being quietist about the correct methodology. This is turn allows for the construction of a taxonomy of moral psychological approaches that allows researchers to efficiently locate the direct source of disagreements. The second chapter examines Lawrence Kohlberg's research program and identifies a particular assumption that guides that program while blocking further progress in the field. That assumption concerns the relation between normative theorizing and descriptive categorization, such that the explanations for why one moral theory is superior to another mirrors stages of moral development. By abandoning this assumption and other assumptions in its local vicinity, researchers could make progress without being bogged down in first-order normative disagreement. The third chapter looks to a recent debate concerning whether neuroscience is normatively significant. Against a standard interpretation of the debate

in Anglophone philosophy that the argument for the normative insignificance of neuroscience is sound, I argue that the critique is only partly successful and not for the reasons commonly recognized. Rather than object to the program of demonstrating the normative significance of neuroscience on normative grounds, we ought to object to the program on descriptive grounds. Each chapter proceeds through arguments rooted in philosophical analysis and reflections on findings in the social and natural sciences – in particular, history, psychology, economics, sociology, anthropology, cognitive science, and neuroscience.

TABLE OF CONTENTS

ACKNOWLEDGMENT ..... **IV**

ABSTRACT ..... **X**

PREFACE ..... **XIV**

DISAGREEMENT ABOUT THE DOMAIN OF MORAL PSYCHOLOGY **1**

1 Introduction **1**

2 What is the subject matter and definition of moral psychology? **3**

2.1 Revision of Original Method **4**

2.2 Popular Conceptions of Moral Psychology **7**

2.3 Three Fundamental Disagreements **10**

3 The domain of moral psychology is the structure of moral cognition **20**

4 The Definition and a Taxonomy of Moral Psychology **23**

5 The Methodological Taxonomy **24**

6 Conclusion **29**

Bibliography **31**

MORAL PSYCHOLOGY, NOT MORALIZED PSYCHOLOGY: REFLECTIONS ON KOHLBERG **33**

1 Introduction **33**

2 The Assumption: Motivations **37**

3 The Assumption: Its Consequences **42**

4 The Assumption: Its Problems **47**

5 The Minimalist Alternative **49**

6 Meta-Theoretical Considerations in Favor of Minimalism **58**

7 Conclusion: Realism and Naturalism in Moral Psychology 60

Bibliography 64

LOCATING THE NORMATIVE INSIGNIFICANCE OF NEUROSCIENCE . 65

1 Introduction 65

2 Summary of Greene's Original Work 66

3 Berker's Dilemma 71

4 Critique of Berker 74

4.1 First Horn Explained 74

4.2 Reply to First Horn 77

4.3 Evaluating the Second Horn 80

5 Greene's Move to Moral Theory 86

6 Possession Argument 91

7 The Modal Objection to the Possession Argument 95

8 21st Century Psychology Cannot Save the Possession Argument 99

9 Conclusion 113

Bibliography 118

PREFACE

This is a collection of essays on moral psychology that critiques the domain and practice of moral psychology by providing an examination of outstanding problems in the field. The title is meant to call to mind Bertrand Russell's famous *The Problems of Philosophy*. A relatively short collection, *The Problems of Philosophy* finds Russell working on a project wherein he tried to say something positive and constructive rather than merely negative about what he took to be some central problems in philosophy (Russell 2001). From this aim, he largely confined himself to epistemological questions concerning knowledge rather than metaphysical questions concerning being. But although I mean to echo both the title and the aim of saying something positive and constructive, my collection of essays does not sidestep the issue of metaphysical disagreement but rather centers it. These essays are related less by unity of argumentation than by unity of concern. The concern is that persistent disagreement in moral psychology is bad for both theoretical and practical reasons.

Theoretically, this kind of persistent disagreement is troubling but also to be expected if we accept a general Sellarsian outline that there is an issue of reconciling the manifest and scientific images of humanity (Sellars 1963). One way of capturing the basic Sellarsian idea is that we have representations of ourselves that are of two kinds: we appear as both persons who exist in the world and act on the basis of intentions and biological creatures with material constitutions that behave in a world of causes, and there is an tension between these two images inasmuch as each image does not easily lend itself translatable in the language of the other. If we accept this overall problematic, then it is

reasonable to expect that moral psychology is going to be of special interest as a domain of inquiry, as suggested by Sellars near the end of his essay: that we are persons with certain ethical standards which can conflict with desires (and, although Sellars does not say this, other kinds of preferences) and to which we may not conform must be reconciled with the scientific image of humanity and human behavior in order for the synoptic response to the problematic to succeed.

Although this collection works in the wake of the Sellarsian problematic, in that a theme concerns coming up with a naturalistically plausible and acceptable account of moral psychology that can do justice to the two images, it is perhaps less optimistic about the proposal to cash out the synoptic vision through ultimately tendentious normative concepts of rights, duties, and community intentions. Rather, we start with the aim of showing that many of the disagreements in moral psychology are in fact not proper to the domain. Once we realize that, we can see some light toward resolution of local problems. The approach is piecemeal and does not propose to solve all the problems in one fell swoop. Rather, we must first diagnose the problems in order to make local improvements that do not rely on the tendentious antecedent normative commitments that drive counterproductive disagreement.

But there are practical concerns as well. Philosophers have an obvious interest in the understanding of the nature, facts about, and possibilities of the human. Natural and social scientists, governments, NGOs, and other individual and institutional change agents have an obvious interest in understanding the causes of human behavior in order to change that behavior. While much work has been completed on the nature of narrowly self-interested, prudential, and

social norm-related behavior, the domain of moral behavior remains underdeveloped, at least compared to the relatively robust predictive success in the other domains in the typology. But we know that moral thinking and behavior exists. We know that it has profound and pervasive impacts on the world and the life prospects of human beings, especially on the life prospects of the least-advantaged persons. Persistent disagreement in moral psychology runs the risk of the development of a constellation of theories that, while empirically adequate relative to antecedent normative commitments, is empirically inadequate with respect to capturing all behavior that is not captured by, for example, rational choice and game-theoretic analysis. The moral, as such, is incalcitrant to such analyses, leaving only rival moral theories to provide guidance in theory choice.

In focusing on the nature of disagreement in the domain of moral psychology, I have come to the conclusion that much of the disagreement concerns issues that, in a certain sense, go beyond the realm of pure psychology. Or rather, the disagreement concerns disagreement, at least in part, over the right way to conduct psychological investigation of the moral. The other part, in my view, is fundamental disagreement about the nature of morality itself. Putting these points together, we can say that the field of moral psychology is riven with methodological and metaphysical disagreement, where "metaphysical" here picks out the nature, if any, of moral reality itself.

Here is a fundamental problem in the field of moral psychology: there is no accepted definition of "moral psychology." That wouldn't be too troubling if there were a more-or-less shared methodology or set of shared methodologies.

Unfortunately, there isn't. But even if there were not a shared definition or more-or-less shared methodology, moral psychology wouldn't be that bad off if there were agreement about the subject-matter of moral psychology. But, again, there is not.

I believe that a plausible explanation of these facts – the facts of disagreement about the subject-matter, definition, and correct methodology of moral psychology— is that those investigating moral psychology deeply disagree about the correct account of morality (they disagree metaethically), about what would be the right or wrong thing to do in some particular circumstance (they disagree normatively), and about what kinds of explanations the moral psychologist should offer, in particular with respect to the role of intuition and theory in explanation (they disagree metaphilosophically). These disagreements are reflected in the nature of their moral psychologizing.

Because of such disagreements, moral psychology is a fractured discipline, if it is a unified discipline at all. The largest division, in my view, is between those working in what I call the humanistic and empirical traditions of moral psychology. Roughly, the humanistic tradition makes use of a wide range of methods and evidential sources to come to particular and determinate claims about moral psychology. In particular, humanists accept as legitimate the methods of conceptual analysis, intuition pumping, and conceptual genealogy and are willing to accept a broad range of evidence, including appeals to philosophical moral theory, emotion, literature, history and the humanities. Empiricists, on the other hand, are restrictive with respect to both method and evidence; in particular, they tend toward accepting only the methods of the

successful natural and social sciences and only the evidence that would counts as "scientific evidence" within those domains.

The humanist and empiricist traditions split on many important problems relevant to the domain of moral psychology. Although not fully articulated and defended only lightly in the collection, there is a broader regulatory ideal of unifying (in principle) moral psychology by putting forward a method that each side – humanist and empiricist – could accept. I call this research program Minimal Moral Psychology. By "minimal moral psychology," I intend to pick out a method of interpretation that guarantees to eliminate as much of the disagreement as possible by deliberately limiting the amount and kind of metaethical, normative and metaphilosophical inputs allowed in moral psychology. This method is a method of ideal interpretation and consists of the application of two principles:

> **Principle 1**: Do not import distinctively moral content into psychological explanations if there are ready-to-hand non-moral psychological tools that can explain the phenomena.

and

> **Principle 2**: "[I]dentify an excess of moral content in psychology by appealing first to what an experienced, honest, subtle, and unoptimistic interpreter might make of human behavior elsewhere" (Williams 1995). And, of course, trim the excess.

Together, the principles, when understood properly and acted upon, help solve the problem of metaethical, normative and metaphilosophical disagreements driving moral psychological disagreement by calling upon the moral

psychologist to cull, if possible, the inputs driving the moral psychological disagreement.

In this collection, I provide three essays that speak to the nature of this methodological and metaphysical disagreement and trying to offer something positive and constructive along the way by offering a way to categorize it, understand at least one of its historical origins, and resolve at least one kind of it.

The first essay provides an understanding of the domain of moral psychology by attending specifically to certain instances of outstanding disagreements and using them to organize thought about the subject matter and definition of the subject. A background guiding principle is that our understanding of the domain of moral psychology should be generous and deflationary enough to capture all the parties in these particular disputes while also being able to separate moral psychology from, say, vision science or branches of social psychology simpliciter – that is, to say something cognitively significant while not also using particular tendentious claims to define away the opposition in these disputes. I argue that there is substantial disagreement about the definition, subject matter, correct methodology of moral psychology. Most of the disagreement is moral (normative), metaethical (or metanormative), and metaphilosophical. We can create a taxonomy of moral psychology along methodological divides. Moral, metaethical, and metaphilosophical disagreement drives much of the between-camp and within-camp disagreement -- especially between those we can identify as "empiricists" and those as "humanists."

The second essay tracks a historical origin of at least one source of persistent disagreement in moral psychology. By focusing on the case of Lawrence Kohlberg, we can come to see more clearly the danger of tying particular descriptive accounts to particular normative accounts, and vice versa. Although many contemporary moral psychologists disavow various parts of his program or its results, it remains the case Kohlberg has exercised a large influence on the field of empirical moral psychology. In particular, Kohlberg assumes that the structures of descriptive moral psychology and normative ethical theorizing will be isomorphic. I argue that this assumption is optional for a moral psychologist. Further, because optional, moral psychologists have freedom to reasonably reject the assumption. Finally, because there also exist compelling metatheoretical considerations against the assumption, moral psychologists have at least some reason to consider an alternate paradigm of moral psychological research centered around the question of how much substantive moral content ought a researcher allow into their program. A background idea here is that moral psychology has become balkanized at least in part because various traditions have built up rival explanations out of the resources of their preferred normative theories. And given that each may be "empirically adequate" relative to the antecedent normative commitments at hand, no tradition has a reason to move from their accounts unless they also move from their normative commitments, which is unlikely for most practitioners. If we allow, though, as seems reasonable to me, that descriptive accounts can constrain normative commitments but that we ought to be wary of normative commitments constraining the descriptive accounts, then it would be

useful to identify normative commitments that are shared across rival traditions in order to assess the empirical adequacy of various descriptive accounts on offer. This is because, in the end, there is no way to do moral psychology but through partial grounding in moral theory.

The third essay concerns the direct interactions between neuroscience, psychology, and normative ethics. One current area of dispute is whether neuroscientific data is normatively significant -- that is, whether we can draw normative conclusions from neuroscientific evidence. Selim Berker and other humanists working in the philosophical tradition argue that we cannot, while Joshua Greene and other empiricists working in the neuroscientific and social scientific traditions have argued that we can. I argue that both sides are partly right and partly wrong. Neuroscientific evidence can serve as partial grounding for normative conclusions in mixed arguments with both normative and empirical premises, contra Berker. However, we should not think, as Greene and others think, that any such arguments will eliminate or select potential candidates for universal morality systems. The debunking will be local -- neuroscientific evidence cannot show us that we ought to be classical utilitarians. This has important implications for research in both the empirical and normative domains. Empiricists should not think that they can defeat their normative rivals merely by means of brain data, but normativists should not think that empirical considerations are wholly irrelevant to figuring out how one ought to live. I allow, in the end, the high-level normative principles often offered as principles of right action may, in principle, be immune to empirical evidence, but mid-level principles must be responsive in order to satisfy a minimal notion of action-

guidance that the overwhelming majority of ethical theories accept as a constraint on adequate ethical theorizing.

This collection of essays attempts to harmonize conceptual analysis, thought experiments, social and natural scientific data to a degree that is respectable in each relative mode of inquiry.

BIBLIOGRAPHY

Russell, B. (2001). *The problems of philosophy*. OUP Oxford.

Sellars, W. (1963). Philosophy and the scientific image of man. *Science, perception and reality*, 2, 35-78.

Williams, B. (1995). *Making sense of humanity: and other philosophical papers 1982-1993*. Cambridge University Press.

DISAGREEMENT ABOUT THE DOMAIN OF MORAL PSYCHOLOGY

1 Introduction

The title of this chapter is "Disagreement about the Domain of Moral Psychology," and the first conclusion of this chapter is that *the domain of moral psychology is the structure of moral cognition*. As such, moral psychology seeks to provide, in part or in whole, an explanation of the structure or some aspect of the structure of moral cognition. There are many aspects of the structure of moral cognition: the psychological and neurophysiological underpinnings of moral and ethical beliefs, judgments, choices, emotions, preferences, motivations, attitudes, and behaviors; the contents of moral and ethical beliefs, judgments, choices, emotions, preferences, motivations, and attitudes; the relations between underpinnings, the relations between contents, and the relations between underpinnings and contents; the relations between the underpinnings and contents of moral cognition and other types of cognition, such as rational choice or social cognition; the presuppositions of different kinds of moral or ethical thinking; the role of the environment in shaping the individual/group and the individual/group shaping the environment; and the evolution and history of the structure of moral cognition. Any work that seeks to provide, in part or in whole, an explanation of the structure of moral cognition of some aspect of the structure of moral cognition counts as a work of moral psychology, according to this view.

The second conclusion of this chapter is that given that we accept the above account of the domain of moral psychology, we should accept a taxonomy

of moral psychology that divides categories along methodological lines. There is no methodology or set of methodologies that is shared by all of those investigating the structure of moral cognition. Dividing by methodology allows us to easily identify methodological disagreement. However, within a particular methodology, there is still substantial disagreement. This within-camp disagreement is largely normative, metaethical and metaphilosophical. We can think of each node in the taxonomy as wrapped in normative, metaethical and metaphilosophical layers. Alternatively, we can think of the taxonomy itself as wrapped in normative, metaethical and metaphilosophical layers. Regardless, normative, metaethical and metaphilosophical disagreement occurs at each node in the taxonomy and explains much within-camp disagreement. Because the normative, metaethical and metaphilosophical disagreement of interest is largely a within-camp phenomenon, it's useful to have a taxonomy that divides along methodological lines.

The two conclusions resolve two outstanding problems in the field of moral psychology: first, "what is the subject matter and definition of moral psychology?," and second, "what are the methods someone could use to examine the subject matter of moral psychology?" Since resolving these two outstanding problems is important for the field of moral psychology, the two conclusions are important for the field of moral psychology.

This paper leaves to the side the question, "Which are the *good* or *right* or *most useful* methods or set of methods someone could use to examine the subject

matter of moral psychology?" It also leaves to the side the question, "Assuming that there is some correct methodology, how do we come to know about it?"

The order of the paper is as follows: first, I will review some literature that suggests that there is disagreement about the definition and subject matter of moral psychology. I will then offer a fuller treatment of my view that the domain of moral psychology is the structure of moral cognition and that moral psychology seeks to provide, in part or in whole, an explanation of the structure or some aspect of the structure of moral cognition. I then provide a sketch of a taxonomy based upon different methods someone could use to explain the aspects of the structure of moral cognition. I explain the taxonomy and argue that it is useful for diagnosing between-camp and within-camp disputes. At the same time, there are some moral psychologists who have mixed methodologies or who occupy more than one camp at a time. The methodological taxonomy can also explain how their work differs from closely related work that does not cross camps.

2 What is the subject matter and definition of moral psychology?

One of the major problems in moral psychology is that there is deep disagreement about what counts as 'moral psychology.' The aim of this section is to say what counts as 'moral psychology' by describing the domain of moral psychology and then defining "moral psychology" in such a way that it offers explanations of target phenomena in the domain of moral psychology.

2.1 Revision of Original Method

In a previous version of this paper that I presented as a talk to the Dissertation Seminar at the University of Pennsylvania, I contrasted two ways of defining "moral psychology." One way to divide the sheep from the goats (moral psychology from something else) is to begin with a top-down definition of "moral psychology." Call this "Fiat Method." An alternative strategy is to observe what people who claim to do moral psychology actually do and to theorize about the relations between these different practices. Call this "Geographical Method." These methods roughly correspond to the difference between conceptual-analytic and social practice accounts of a particular domain of inquiry. I argued that the Geographical Method was a better way of defining "moral psychology" because I thought the method was less tendentious and less open to disagreement than the Fiat Method. I also thought the Geographical Method was useful inasmuch as it points our attention toward the actual practices of people who claim to do moral psychology, and such facts, I claimed, were pretty important in thinking about moral psychology.

This distinction between these two methods is ultimately futile, although there is something important that the distinction aims to capture. What is ultimately futile about the distinction is that the Geographical Method is, at bottom, itself a Fiat Method, inasmuch as all strategies for definition rely upon operationalization of terms. All definitions are "top-down" in this sense, and there is no meaningful distinction between Fiat and Geographical Methods in how I previously described the methods. That said, there is an important

difference between defining "moral psychology" in terms of one's preferred approach to moral psychology and defining "moral psychology" in terms of practices related to the domain of moral psychology. The former is too exclusive, depending on what one's preferred approach to moral psychology is.[1] But the latter doesn't strike me as too exclusive or inclusive: it hits a Goldilocks standard in terms of scope by limiting the account to observable behavior and practice.

That said, my original formulation was incomplete and open to misunderstandings. In the original formulation, I attempted to sidestep the issue of specifying the target domain of moral psychology. I thought that I could sidestep the issue since I could identify the target domain downstream; in particular, I thought that by appealing to the practices of people who claim to do moral psychology, I could then, from those practices, identify the subject-matter of moral psychology and that such identification would be protected from standard objections to domain specification. I was wrong.

There are a couple of objections that arise in response to a proposal like that of my original approach, and most of these objections are rooted in the fact that such an approach doesn't specify individually necessary and jointly sufficient conditions for membership that would exclude prima facie absurdities. So, even though objections based on rhetorical questions such as, "Is my bowling

[1] For me to define "moral psychology" in terms of my preferred approach to moral psychology would make me almost certainly have to rule out many practices that I had wanted to call "moral psychology."

ball moral psychology?" are, strictly speaking, non-sequiturs,[2] there are related issues. Namely, my previous account, the one not tied to a subject-matter, didn't suitably restrict moral psychology because it allowed that moral psychology is (i) what people who claim to do morally psychology actually do and (ii) the relations between those practices. But, moral psychologists do lots of things – they brush their teeth, they listen to trap music, they study chemistry, and so on. So it is insufficient to say that moral psychology is what people who claim to do moral psychology actually do, for people who claim to do moral psychology actually brush their teeth, but no one thinks that brushing your teeth is moral psychology.[3]

Because of this, I have come to realize that I must say something about the subject-matter of moral psychology in order to fruitfully address the question, "what is moral psychology?" We need to restrict the range of activities that moral psychologists actually do that are relevant for picking out moral psychology as a field. We could try the strategy of saying that moral psychology is the stuff people who claim to do moral psychology do when they claim to do moral psychology. This again allows objections from deviant cases and

[2] Because the point was never "moral psychology is whatever people who claim to do moral psychology claim to be." Basically, the "bowling ball objection" only would work on the condition that my account was meant to say, "if a person who claims to do moral psychology claims to be some X, then moral psychology is also that X," and if it were true that there were some individual who both claimed to do moral psychology and to be a bowling ball. I assume such a person doesn't exist. The point was "moral psychology is whatever people who claim to do moral psychology actually do."

[3] I owe part of this line of questioning to Daniel J. Singer and Andrew McAninch.

absurdities.[4] I see no way to avoid these objections but to specify the domain and subject matter of moral psychology. Importantly, though, this doesn't mean that I've given up on all of the original approach. Rather, we must integrate the practices of people who claim to do moral psychology with an account of the domain and subject matter of moral psychology into our definition of "moral psychology," and we'll have to do so in a more or less holistic way, even though the subject-matter of moral psychology is given some explanatory priority in marking out what counts as moral psychology.

2.2 Popular Conceptions of Moral Psychology

What is the subject matter and definition of moral psychology? Thankfully, like many questions of this sort, we do not have to start at the barest of philosophical intuitions and work our way up or deduce our way down from there. Instead, there is a history of practices aimed at providing explanations in the target domain of moral psychology. And sometimes the practitioners have provided definitions of moral psychology that we can now evaluate. In what is to follow, I will present a series of definitions of moral psychology that have popped up among those aiming at providing explanations in the target domain of moral psychology. This series is not exhaustive, but it is meant to illustrate

[4] For example, we could have a deviant case where a person claim to do moral psychology performs some behavior and calls that behavior moral psychology, but we simply wouldn't want to say that it's moral psychology. But the deviant who claims to do moral psychology and claims that snapping his fingers 13 times before leaving the room counts as "moral psychology" isn't doing moral psychology. He's doing some other, bizarre thing.

important continuities and discontinuities between rival conceptions of moral psychology.

In the introduction to their collection *Moral Psychology: Historical and Contemporary Readings* (2010), Thomas Nadelhoffer, Eddy Nahmias and Shaun Nichols write:

> Moral psychology is the field that addresses these and related issues – it is the study of the way humans think about morality, make moral judgments, and behave in moral situations. While the immediate goal of the field is to understand moral cognition and behavior, the inquiry also has possible implications for how we should make moral judgments and how we should behave. Even though we cannot move directly from data concerning how we actually do think and behave to theories about how we ought to think and behave, by exploring morality in an interdisciplinary way, moral psychologists are, at a minimum, able to place empirical constraints on normative theorizing. Moral psychology thus involves the intersection of philosophy and empirical sciences ranging from evolutionary biology and game theory to neuroscience and social psychology. (p. 1)

"These issues" that moral psychology addresses as a field are

> What is it about human beings that enables (or compels) us to engage in such complicated moral thought and behavior [e.g., non-reciprocal altruism and moral debate]? What biological and psychological capacities

> underlie our moral judgments? What drives us to help those in need? What enables us to follow moral norms and to be responsible for transgressing them? (p. 1)

Nadelhoffer, Nahmias and Nichols capture important features of moral psychology in their definition, even if we may wonder about some of their particular details of their account.[5] Taken in a suitably broad way, it is undeniably true that moral psychology has the subject matter of "the way humans think about morality, make moral judgments, and behave in moral situations." A complaint about this way of talking is that it is far too coarse-grained to be informative beyond the obvious platitude that moral psychology is about morality and psychology (allowing that psychology is about thinking and behavior). But this complaint is lessened in view of the questions that moral psychology addresses as a field. To take one example, what is it about us human beings that enables or compels us to engage in moral debate? This is an example that begins to flesh out the subject matter of moral psychology in a way that says something non-trivial and important about moral psychology. We can investigate the question from cognitive scientific or sociological points of views, and what we have to investigate is a particular aspect of the structure of moral cognition.

[5] For example, if moral psychology has "possible implications" for ethics and ethical theory, then the best they can say is that, at a minimum, moral psychologists only possibly are able to place empirical constraints on normative theorizing. And the claim that "moral psychology thus involves the intersection of philosophy and empirical sciences" doesn't follow from "by exploring morality in an interdisciplinary way, moral psychologists are, at a minimum, able to place empirical constraints on normative theorizing." But I set these issues to the side for the sake of discussion.

2.3 Three Fundamental Disagreements

Disagreement 1: Is the interface of empirical psychology and normative ethics a topic in moral psychology?

Another important feature of Nadelhoffer, Nahmias and Nichols's definition is that it incorporates normative ethical theorizing (or at least implications for normative ethical theorizing in the form of "constraints") into the field of moral psychology. Another way to put the point is that, according to them, the interface of what we might call "empirical moral psychology" and normative ethical theorizing is itself part of the subject matter of moral psychology.

This view has support among other contemporary moral psychologists. So, for example, John Doris writes in the introduction to *The Moral Psychology Handbook* (2010) that for moral psychology, times lately have been both interesting and good: research at the intersection of human mentation and human morality is flourishing as never before" and that "the discipline of moral psychology is, as the name intimates, a hybrid inquiry, informed both by ethical theory and psychological fact" (p. 1). Or consider the account offered by Doris and Stephen Stich in the *Stanford Encyclopedia of Philosophy* article "Moral Psychology: Empirical Approaches" (2017):

> Moral psychology investigates human functioning in moral contexts, and asks how these results may impact debate in ethical theory. This work is necessarily interdisciplinary, drawing on both the empirical resources of

the human sciences and the conceptual resources of philosophical ethics." (par. 1).

These all suggest that "moral psychology" is not (or *not only*) what contemporary psychologists in departments of psychology at universities do when working on explanations of the causal and computational structure of moral cognition.

But not all investigators agree that moral psychology includes the interface of empirical moral psychology and normative ethical theorizing. For example, Regina Rini (2015) argues that moral psychology is about the causal and computational structures of the human moral faculty.[6] This suggests to me that she thinks that the interface of moral psychology and normative ethical theorizing is not itself part of moral psychology. This investigation of the interface would belong to a branch of philosophy, perhaps moral philosophy or metaphilosophy. And we can draw upon many examples of moral philosophers – from Kant (1998) to many contemporary Anglo-American or "analytic" philosophers – who think that empirically-oriented moral psychology (one of the two aspects of moral psychology in the Nadelhoffer, Nahmias, and Nichols and Doris, Doris and Stich lines) is largely or completely irrelevant to the practice of normative ethical theorizing. For example, Selim Berker (2009) argues in "The Normative Insignificance of Neuroscience" that

[6] It is true that Rini qualifies her statements so that it is about "empirical moral psychology," which leaves open the possibility that there is a non-empirical moral psychology. But, later in her article, she slides between empirical psychology and psychology when making her claims about the causal structure of the human moral faculty.

> [E]ither attempts to derive normative implications from these neuroscientific results rely on a shoddy inference, or they appeal to substantive normative intuitions (usually about what sorts of features are or are not morally relevant) that render the neuroscientific results irrelevant to the overall argument. (p. 294)

What is animating Berker's position here is that the interface of empirical moral psychology – at least neuroscientific versions of empirical moral psychology – and normative ethical theorizing is a subject in philosophy and not in moral psychology.[7]

So we have disagreement between those investigating moral psychology. Some think that moral psychology includes the interface of empirical moral psychology and normative theorizing, and others disagree.

Disagreement 2: Must moral psychology make use of the resources of philosophical ethics?

But this is not the only disagreement. Again, the Nadelhoffer/Nahmias/Nichols, Doris, and Doris/Stich accounts claim that moral psychology is an interdisciplinary field and that one of the relevant disciplines is philosophical ethics itself. But there are some people working on moral psychology who do not make use of the resources of philosophical ethics and who would typically eschew such resources. The reasons for avoiding the

[7] I interpret this quote as saying that attempts to lay the groundwork for the interface fall prey to bad epistemological practice. The interface is a proper subject of logic, philosophy of science, and epistemology. The proper (philosophical) view is that there's no such interface.

resources of philosophical ethics are varied: perhaps the resources aren't vouchsafed in the right way, or perhaps the resources are not properly "empirical," or perhaps the resources can't be falsified or whatever the standard (and that the resources can't be falsified or whatever constitutes an obstacle to making use of the resources).

For example, Cristina Bicchieri does work on what is recognizably "empirical moral psychology," but she does not make use of the resources of philosophical ethics. Bicchieri's theory of social norms is a system of classification of characteristic motivations of collective patterns of behavior (2006, 2016). Using traditional concepts associated with game-theoretic analysis and rational choice theory (such as 'preference' and 'belief,' she distinguishes between customs, descriptive norms, social norms, and moral norms. According to Bicchieri, moral norms are a subset of personal norms, and personal norms are marked by an unconditional preference to act in accord with the norm.[8] This is in contrast to unilateral and multilateral descriptive norms and social norms, where the preference to act in accordance with the norm is conditional on empirical expectations that relevant others act in accordance with the norm or conditional on empirical expectations plus normative expectations. Normative expectations are beliefs about other people's beliefs about what you should or what should be done. Behavior in compliance with moral norms is grounded in personal normative belief, a first-order belief of the form "I think I/you/they/us should do X."

[8] Habits, customs and moral norms are all instances of personal norms, according to Bicchieri's 2006 account.

Bicchieri's 2006 treatment of moral norms and moral norm compliance is the least developed treatment of any of the collective patterns of behavior that she examines. But, in laying out her partial theory of moral norms, Bicchieri does not seem to call upon the resources of philosophical ethics. Now, it's true that there is a tradition in philosophical ethics that maintains that moral norms issue unconditional demands, and there is a philosophical ethical tradition that says that when a person makes a moral judgment, then, *ceteris paribus*, that person is motivated to act on the judgment regardless of what she perceives others to think of her moral judgment. But Bicchieri makes the distinction between moral norms and, say, social norms not on the basis of philosophical ethics but rather on the basis of the belief/preference model from rational choice theory and the concept of interdependent choice from game theory. And there is no part of her account that implies the motivational internalism from the philosophical ethical tradition, as it remains possible (and in many cases of sufficient social pressure, likely) that individuals with not act in order with the moral belief that they happen to hold with respect to some behavioral rule. This is why, I think, this account has difficulty separating habits/customs from moral norms. So, for example, Bicchieri (2006) writes

> Condition 2 (the *conditional preference condition*) marks an important distinction between social and personal norms, whether they are habits or have moral force. Take the habit of brushing my teeth every morning. I find it sanitary, and I like the taste of mint toothpaste. Even if I came to realize that most people stopped brushing their teeth, I would continue to

> do so, because I have independent reasons for doing it. It is likewise with moral norms: I have good, independent reasons to avoid killing people I deeply dislike. Even if I were to find myself in a Hobbesian state of nature, without rules or rights, I would still feel repugnance and anguish at the idea of taking a life. With this I do not mean to suggest that moral norms are a world apart from other rules. Instead, by their very nature, moral norms demand (at least in principle) an unconditional commitment. (20)

Both habits and moral norms are motivated by unconditional commitments (that is, unconditional on what I expect others to do or believe), and we can express this fact by noting that habits and moral norms are each grounded in (good) independent reasons.[9] Still, Bicchieri says that our motivational profile toward the collective behavior or rule of behavior determines whether we are dealing with moral norms or social norms, not our stereotypical justifications of the behavior that we would attempt to give. But most moral philosophers, especially working in the wake of 20th century metaethics that try to separate morality from other normative domains such as prudence or etiquette, and, although this is specuation, most ordinary non-philosophers explicitly avow that habits or

[9] I think that there is a problem here with the introduction of "good" before "independent reasons to avoid killing people I deeply dislike." It makes sense to say that I have good, independent reason to brush my teeth. Since I value my overall health, it is instrumentally rational and good for me to brush my teeth. But "good" doesn't seem to function in the same way in the imagined moral case. It may not be instrumentally rational of me not to kill people I dislike in a Hobbesian environment. So I don't know what Bicchieri means here. "Good" in the case of habit clearly picks out "good for me," and in some sense over and above "good for me in the sense that I think the reasons are morally good." People sometimes express that that have good moral reason to do such-and-such, even though doing such-and-such would be "bad" for them on all but the most ascetic readings of "bad for you."

matters of prudence are different than moral norms, even though both demand "unconditional commitment" in the sense that Bicchieri employs the term.[10] If Bicchieri were drawing upon the resources of philosophical ethics, then she should be able to cleanly separate habits from moral norms in line with the philosophical tradition and ordinary understanding. Therefore we can conclude that she is not drawing upon the resources of philosophical ethics.[11]

So far, we have two disputes about the subject matter and definition of moral psychology. First, some think that moral psychology includes the interface of empirical moral psychology and normative theorizing, and others disagree. Second, some think that moral psychology, perhaps regardless of interface issues, requires the conceptual resources of philosophical ethics, and other disagree. I want to turn now to one final point of dispute.

Dispute 3: Should moral psychology be examined in an empirical or humanistic way?

This point of dispute is a growing issue, and I think that it highlights something important about moral psychology. Remember that we are examining accounts that maintain that moral psychology is a hybrid theory of ethical theory and psychological fact. The "psychological facts" here correspond to the facts as given by empirical investigations into moral thinking and behavior. But there are some people who think that there is a distinctly *philosophical moral psychology* that

[10] Or at least our (moral philosophers' and perhaps many of the folks') concepts "habit" and "moral norm" are clearly different.
[11] Of course, this assumes that rational choice theories are not examples of philosophical ethics. I would be willing to consider treating the theories as ethical theories; in such a case, I would need an alternative example of my point here.

can be contrasted with *empirical moral psychology*, and they think that philosophical moral psychology is *superior* to much empirical moral psychology (or is a necessary supplement to empirical moral psychology in order to curb its excesses). The question here is whether moral psychology belongs to the empirical sciences. Some say yes, and others disagree.

A couple of examples here should suffice. Carla Bagnoli (2011) identifies the dispute between empirically-oriented moral psychology – of the sort practiced by Stich, Doris and other people who are members of or who are influenced by the Moral Psychology Research Group – and what I've called above "philosophical moral psychology. She writes:

> In contrast to this empirical approach to philosophy and psychology, others argue that both disciplines are autonomous with respect to the cognitive sciences. [Here she cites Bernard Williams and R. Jay Wallace as exemplars] Of course, empirical findings may be indirectly relevant to philosophical arguments. . . Even if they recognize that a dialogue with the empirical sciences is inevitable and rewarding, these philosophers argue that it is misleading to think of the activity of philosophy as modeled on the empirical sciences. At issue, then, is not science, but 'scientism', [sic] or the philosophical view that assimilates philosophy to science and borrows its methods. (p 13)

Bagnoli goes on to discuss how the issue turns on the issue of *moral motivation* and how many analyses of moral motivation depend, in part, on *a priori* methods

of analysis. We see this sort of issue pop up when moral philosophers talk about empirical treatments of, say, altruism and altruistic behavior. The empirical psychologist or philosopher offers some operationalized sense of "altruism" and "altruistic behavior" that is testable and allows for measurement. The moral philosopher says, "You've missed important part P of altruism (typically, a characteristic of the motivation). Your operationalization is bad, you've missed the target phenomena, and your results don't show what you claim that the show."

In order to demonstrate that this is a real phenomena, one could look at the vast literature comprising Humean, Kantian, Aristotelian, and Thomist theories of moral motivation and see how they rely upon *a priori* and normative methods that sharply distinguish the moral from the non-moral. For a particular case in point, we could look at Lawrence Blum's criticisms of Shaun Nichols's sentimental rules account of moral cognition. Blum (2011) lays out his complaints at length:

> I will argue that Nichols's view suffers from several deficiencies: (1) It operates with an impoverished view of the altruistic emotions (empathy, sympathy, concern, compassion, etc.) as mere short-term, affective states of mind, lacking any essential connection to intentionality, perception, cognition, and expressiveness. (2) He fails to keep in focus the moral distinction between two very different kinds of emotional response to the distress and suffering of others – other-directed, altruistic emotions that have moral value, and self-directed emotional responses, such as personal

> distress, that do not. (3) Nichols is correct to see morality as requiring affectivity, the capability of emotional response to others; but his incorrect view of altruistic emotions (and of emotions in general) leads him to misstate the connection between morality and emotion. (4) Nichols fails to recognize Schopenhauer's form of anti-rationalism as distinct from Humean sentimentalism; some of his arguments presented to support the latter instead lend support to the former. (5) Finally, while agreeing that moral philosophy is strengthened by knowledge of empirical psychology, I suggest that the foregoing failures of Nichols's argument are partly due to his misuse of particular empirical results and findings, his being over-enamored of empirical psychology, and possibly to a weakened commitment to the distinctive contribution that the humanistic methods of philosophy make to our understanding of the moral enterprise. (p. 171)

Blum's disagreement with Nichols's is as complete as I could wish for in an example, and Blum's criticisms of Nichols perfectly encapsulate the divide between empirical and humanistic or philosophical modes of moral psychology. Blum argues that Nichols mischaracterizes altruism, that Nichols too readily relies on empirical rather than analytic accounts of altruistic motivation, that Nichols has a generally impoverished view of emotions due to ready reliance on empirical psychology, that Nichols fails to respect the history of philosophical theory (especially philosophical theory with respect to moral psychology, and that Nichols's failures are related to his metaphilosophical attitudes. This is not to say that I endorse Blum's criticisms, nor is it to say that I reject them. The

criticisms, as such, are orthogonal to the point that I'm trying to make. The point that I'm trying to make is that Blum and Nichols substantially disagree about the correct methodology and the role of the experimental method and empirical data in choosing between rival accounts in moral psychology.

So, to summarize, there is substantial disagreement about whether the interface of empirical moral psychology and normative ethical theorizing is a subject in moral psychology, there is disagreement about whether doing moral psychology requires taking on the conceptual resources of normative ethical theory, and there is disagreement about whether we should investigate moral psychology with the tools of experimental psychology, of traditional philosophical analysis, or of some combination thereof. These disagreements amount to disagreements about the subject matter, definition and methodology of moral psychology. If that's true, then I think that we should look for an understanding of moral psychology that locates these disputes as within the field of moral psychology.

3 The domain of moral psychology is the structure of moral cognition

As I said previously, I think that one way to get traction in thinking about the definition is to think about the domain of moral psychology. I've also said that the domain of moral psychology is the structure of moral cognition. And I've said that the structure of moral cognition has the following aspects: the psychological and neurophysiological underpinnings of moral and ethical beliefs, judgments, choices, emotions, preferences, motivations, attitudes, and

behaviors; the contents of moral and ethical beliefs, judgments, choices, emotions, preferences, motivations, and attitudes; the relations between underpinnings, the relations between contents, and the relations between underpinnings and contents; the relations between the underpinnings and contents of moral cognition and other types of cognition, such as rational choice or social cognition; the presuppositions of different kinds of moral or ethical thinking; the role of the environment in shaping the moral and ethical beliefs of the individual/group and the individual/group, from those moral and ethical beliefs, shaping the environment; and the evolution and history of the structure of moral cognition.

This understanding of the domain of moral psychology is, I believe, consistent with what how the people I've talked about so far think about the domain of moral psychology. Granted, I have specified the domain in greater detail than merely talking about moral thinking and more behavior. I have broken down "thinking" into components like "beliefs," "judgments," "choices," "emotions," "preferences," "motivations," and "attitudes." I have separated the discussion of the cognitive and neurophysiological underpinnings of these acts of moral thinking from their contents, and I've deliberately left "underpinnings" and "contents" open so as to be able to capture substantive disputes about what counts as an "underpinning" or as a "content." My characterization of the structure of moral cognition is in principle consistent with both sides of the three disagreements that I've listed before as fundamental disagreements. The structure of moral cognition is the target phenomena that researchers in moral

psychology are after, and many of the people listed above have some substantial disagreement about some aspect of the structure of moral cognition. So, for example, debates about where to locate the interface of empirical moral psychology and normative ethical theorizing take different aspects of the structure of moral cognition as salient and then advance substantive positions about those aspects. The salient aspect, for one theorist, could be the relation between the underpinnings and the contents of moral psychology; for another, it could be about merely the contents of moral psychology. It doesn't really matter for our purposes whether we think that the interface is a subject in empirical moral psychology or normative ethical theorizing – if it's part of empirical moral psychology, then it's a part of the structure of moral cognition, and if it's part of normative ethical theorizing, then it's also a part of the structure of moral cognition (in virtue of the fact that moral psychology seeks to explain the contents of particular moral judgments as well as specifying the casual and computational structure of the human moral faculty).

If this characterization of the structure of moral cognition has a "heads, I win, and tails, you lose" quality to it, do keep in mind that the characterization rules out a lot as possibly being the subject matter of moral psychology. Set theory is not part of the subject matter of moral psychology nor is non-organic chemistry nor is brushing your teeth nor is social norm theory, except inasmuch as these items can become the content of particular moral judgments. There is a universe of not-moral psychology. With that said, it's not an objection to my account of the structure of moral cognition that lots of people talking about

moral psychology are doing moral psychology when they do the sorts of things I was describing them as doing. It's a feature, and not a bug, of the system that those I've mentioned before are all doing moral psychology.

4 The Definition and a Taxonomy of Moral Psychology

But, from this first approximation as to the structure of moral cognition, we get a principled way to define "moral psychology." The definition is that moral psychology seeks to provide, in part or in whole, an explanation of the structure or some aspect of the structure of moral cognition. This definition is principled because we have defined a target class of phenomena that is of cognitive significance. We want to know about those aspects of the structure of moral cognition, and moral psychology is the field that attempts to provide partial or complete explanations of aspects of or all of the structure of moral cognition. If you are not trying to provide partial or complete explanations of aspects of or all of the structure of moral cognition, then you are not doing moral psychology. Alternatively, we do not call you a "moral psychologist."

From the definition of "moral psychology," we could devise a taxonomy that breaks down according to which aspect or aspects of the structure of moral cognition that the theorist is trying to explain. I think that this results in a needlessly complicated schema, and the needless complication gets in the way of any true utility that the schema could provide. Rather, all that we need for the purpose of categorization are methodologies. We assume in the taxonomy that we have already antecedently addressed which aspect of the structure of moral

cognition we want to investigate. Then, we can ask, "What are the methodologies that we could use to investigate this subject matter." Here is where attention to the practices of extant moral psychology because useful again. Although explanatory priority is given to the structure of moral cognition, we cannot ignore the practices of moral psychology if we wish to diagnose primary areas of dispute. We can provide a methodological division according to different styles of explanation, assuming that there are multiple styles and that causal explanations are not the only sorts of explanation. Then, within a methodological division, we can identify within-camp disagreement by invoking substantive normative, metaethical and metaphilosophical positions. I want to diagnose primary points of methodological and substantive dispute, and any representation that allows such a diagnosis and that is clear and perspicuous is acceptable from my perspective.

5 The Methodological Taxonomy

Given that moral psychology is the field that seeks to provide, in part or in whole, an explanation of the structure or some aspect of the structure of moral cognition, then it makes sense to distinguish among kinds of moral psychology by the characteristic types of explanation that the kinds would give.

In my taxonomic sketch, we can distinguish between humanistic, empirical and theological types of explanation. I've already gone part way toward distinguishing between humanistic and empirical types of explanation. Empirical types of explanation rely upon the scientific method generally and the

methods of the successful natural and social science more specifically to provide explanations of whatever aspect of the structure of moral cognition is under consider. Many people grant that the aspect of the structure of moral cognition concerning the psychological and neurophysiological underpinnings of aspects of moral thinking and behavior either require or are enhanced by means of empirical explanations.[12]

But some people don't grant that theological explanations count as genuine explanations, or they don't grant that theological explanations are autonomous from humanistic explanations more generally. To the first, I reply that positing a moral sense implanted by a deity is a potential explanation. It's not a very good explanation, but, if it were true that a deity implanted a moral sense in humans, then you could potentially explain some aspects of the structure of moral cognition in terms of your proposed theology. To the second, I reply that humanistic explanations are different in kind from theological explanations. Let me explain by talking more about what I think constitute humanistic explanations.

I think that humanistic explanation come in three primary modes: normative, phenomenological and historical/genealogical. These methods try to explain aspects of the structure of moral cognition, but they don't do so through the methods of the successful natural and social sciences or through the methods peculiar to particular theological commitments. Humanistic modes of

[12] The neurophysiological *must* be carried out in empirical fashion. There is no humanistic or theological tradition of explanation that even begins to account for such underpinnings.

explanation are pitched at the level of interpretations of the human being as beings in the natural world. There is a fundamental explanatory difference between positing a faculty of pure practical reason (which would count as "humanistic" under my schema) and positing a moral analogue of a *sensus divinitatis* (which obviously would count as "theological").[13]

Under the humanistic tradition, which I've previously identified with Blum and which Bagnoli identified with Williams, Wallace and the Humean, Kantian, Aristotelian and Scholastic theories of moral motivation, we can identify people who are attempting to explain aspects of the structure of moral cognition by reference to explicit normative theories ("normative"), by reference to the phenomenology of lived moral experience ("philosophical" and "literary" versions – the distinction here is a genre distinction, but it's also related to the explicitly articulated phenomenological theory), and by reference to historical-genealogical accounts that seek to vindicate or deflate confidence in held belief [for example, (Nietzsche 1989), (Foucault 1977) and Williams's (2002) on the multifarious uses of genealogical method).

Against the theological and humanistic traditions, we identify the empirical tradition, and we subdivide according to the venerable distinction in the philosophy of social science between methodological individualism, holism and mixed approaches. Individualist approaches seek to explain collective patterns of behavior by reducing the phenomena to patterns of interlocking individual behavior within certain boundary constraints. Holist approaches seek

[13] I thank Devin Curry for this point.

to explain individual behavior in terms of social entities, institutions, or phenomena that are in some sense *more* than the aggregate of individual choice. Mixed methods mix and match according to the particular subject matter.

What is useful about this taxonomy is that it allows us to identify points of dispute between, say, Thomists, Blum and Nichols. We can say that they are all doing moral psychology, but they disagree with respect to method.

Importantly, the schema also allows us to identify when people are mixing methods. So, for example, Duke Naturalists make simultaneous use of historical-genealogical and empirical methods. We can track that Wong (2009) and others are doing this, and we can separate them from folks who are merely doing one or the other style of explanation. Or, for another example, many Christian moral psychologists of late antiquity combined literary, historical-genealogical, and theological methods {for example, see (Augustine 1998). Again, the schema tells us about possible methodologies; it doesn't tell us which methodology is right or which methods are the good or reliable or useful methods. I will say that I think that it is a function of one's normative, metaethical and metaphilosophical commitments whether one is likely to mix and match methodology, although I will not explicitly argue for that here. It is enough to say that mixed methodologies don't need any special place within the taxonomy, for we have all the components necessary for explaining in what ways a particular mixed methodology happens to be mixed.

But still, there is within-camp disagreement that is also important to account for. The rings or layers of normative, metaethical and metaphilosophical commitment allow us to do just that. We can identify points of within-camp dispute – such as the dispute between Joshua Greene (2008) and John Mikhail (2008) at the level of individualist empirical moral psychology – as largely arising out of normative and metaethical disagreement. They disagree, for example, about whether it would be right to push the fat man in the bridge version of the trolley problem. They also disagree about what the correct account of morality is more generally. But they mostly agree on the correct methodology for the investigation of the psychological and neurophysiological underpinnings of moral thinking – they are both using the methods of cognitive science.

To add some further content to the bare sketch: John Doris (2002) and Gilbert Harman (1999) argue that social psychological research shows that there are no robust character traits as assumed and required by contemporary virtue ethical theories, Sharon Street (2006) argues that Darwinian evolution undermines claims that our psychology is adapted to perceive robustly mind-independent moral facts, and Joshua Greene and many others (2001) argue that there is compelling psychological and neuroscientific evidence that characteristically utilitarian judgments are reason-based (in some interesting way) while characteristically deontological and virtue ethical judgments are emotion-based (in some interesting way).

Of course, there are others working at the same level in the taxonomy and who dispute these claims. So, Rachana Kamtekar (2004), Daniel Russell (2009),

Nancy Snow (2010) and many others argue that Doris and Harman's complaints are overblown and that character traits have respectable scientific credentials and that social psychology is *consistent* with normative virtue ethics. David Copp (2008) and others have argued that Darwinian evolution poses no problem for their versions of naturalistic moral realism. And John Mikhail (2011), Fiery Cushman (2013) and others have argued that characteristically deontological judgments are wholly or partly constituted by cognitive appraisals.

These disputes are largely normative, metaethical and metaphilosophical. Our taxonomy should allow space for these disagreements and account for them. This is not a full account of the ways in which normative, metaethical and metaphilosophical commitments drive within-camp moral psychological disagreement, but it is an introduction. Fuller articulation of this view will depend on the fuller articulation of a method of separating normative and non-normative content so that we can be in a position to *reliably identify* points of normative, metaethical and metaphilosophical dispute. However, we have a general sense of what we're up to, and I see no in principle objection to this line of inquiry.

6 Conclusion

I have offered an account of the domain of moral psychology that identifies it as the structure of moral cognition. I spent quite a deal of time explaining how there is disagreement about the definition of "moral psychology." I argued that the identification of the domain of moral psychology

with the structure of moral cognition was principled and consistent with prevailing attitudes. I then argued that the identification gives us a principled basis for defining moral psychology as the field that seeks to provide, in part or in whole, an explanation of the structure or some aspect of the structure of moral cognition. I then gave a very brief taxonomic sketch about alternative methods of explanation of aspects of the structure of moral cognition.

What we need to develop going forward is a more fully articulated account that shows how the methodological taxonomy interlocks the aspects of the structure of moral cognition such that certain aspects of the structure of moral cognition are more amenable to certain styles of explanation. For now, it suffices to say that I have specified the domain of moral psychology, provided a working definition for "moral psychology" on the basis of that domain specification, and presented a first-pass taxonomy of positions in moral psychology by dividing along methodological lines. Showing how the normative, metaethical and metaphilosophical layers work in the picture require more work and is outside the scope of the present discussion.

Bibliography

Augustine, S. (1998). *Augustine: The City of God Against the Pagans*. Cambridge University Press.

Bagnoli, C. (Ed.). (2011). "Introduction." *Morality and the Emotions*. Oxford University Press.

Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs*, 37(4), 293-329.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Blum, L. "Empathy and Empirical Psychology: A Critique of Shaun Nichols," in *Morality and the Emotions* (ed. Carla Bagnoli) (Oxford University Press, 2011): 170-193

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, 17(3), 273-292.

Copp, D. (2008). Darwinian skepticism about moral realism. *Philosophical Issues*, 18(1), 186-206.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge University Press.

Doris, J. M., & Moral Psychology Research Group. (2010). *The moral psychology handbook.* OUP Oxford.

Doris, John, Stich, Stephen, Phillips, Jonathan and Walmsley, Lachlan, "Moral Psychology: Empirical Approaches", The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/moral-psych-emp/>.

Foucault, M. (1977). "Nietzsche, Genealogy, History." In *Language, Counter-Memory, Practice: Selected Essays and Interviews* (ed. D. F. Bouchard). Cornell University Press.

Greene, J. (2008). The secret joke of Kant's soul. Sinnott-Armstrong W, ed. *Moral psychology, vol. 3: the neuroscience of morality: emotion, disease, and development*.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108.

Harman, G. (1999, January). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. In *Proceedings of the Aristotelian society* (pp. 315-331). Aristotelian Society.

Kamtekar, R. (2004). Situationism and virtue ethics on the content of our character. *Ethics*, 114(3), 458-491.

Kant, I. (1998). *Groundwork of the Metaphysics of Morals*. 1785.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

Mikhail, J. (2008). 2.1 Moral Cognition and Computational Theory. *Moral Psychology: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, 3, 81.

Nadelhoffer, T., Nahmias, E., & Nichols, S. (Eds.). (2010). *Moral psychology: historical and contemporary readings*. John Wiley & Sons

Nietzsche, F., Clark, M., & Swensen, A. J. (1998). *On the genealogy of morality*. Hackett Publishing.

Rini, R. A. (2015). Psychology and the aims of normative ethics. *Handbook of Neuroethics*, 149-168.

Russell, D. C. (2009). *Practical intelligence and the virtues*. Oxford University Press.

Snow, N. E. (2010). *Virtue as social intelligence: An empirically grounded theory*. Routledge.

Street, S. (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies*, 127(1), 109-166.

Williams, B. A. O. (2002). *Truth & truthfulness: An essay in genealogy*. Princeton University Press.

Wong, D. B. (2009). *Natural moralities: A defense of pluralistic relativism*. Oxford University Press.

MORAL PSYCHOLOGY, NOT MORALIZED PSYCHOLOGY: REFLECTIONS ON KOHLBERG

1 Introduction

Our psychological theory as to why moral development is upward and sequential is broadly the same as our *philosophical* justification for claiming that a higher stage is more adequate or more moral than a lower stage.

-Lawrence Kohlberg (1971)[14]

The fit between the special psychological conception and the demands of morality enables us to see that this piece of psychology is itself a moral conception, and one that shares notably doubtful features of that particular morality itself.

-Bernard Williams (1995a) [15]

Lawrence Kohlberg jumpstarted the modern experimental turn in moral psychology.[16] And, even though many empirically-oriented moral psychologists

[14] P. 180-81.
[15] P. 74.
[16] I stress the "modern" here because there is an experimentalist tradition in moral psychology dating back to Aristotle and running through to Hume and the sentimentalist traditions, to Spencer and the evolutionary traditions, and to Dewey and the classical pragmatist tradition. In point of fact, Kohlberg thinks of himself as traveling in Dewey's wake, armed with certain methods adapted from Piaget (1971, p. 154). But, given that Kohlberg's work on morality is probably more read than Dewey's at this point, I see nothing especially problematic about my ascription. Regardless, I take Kohlberg's work – and his assumption about the relation of descriptive and normative practices in moral psychology – to be *symptomatic* of a larger trend. Given that Kohlberg still serves as a shared point of reference for those working in moral psychology, I hope that by starting with him we can recognize a strain of thinking that continues to today.

now believe that Kohlberg's theory was wrong in one way or many, his influence stills extends throughout the community.[17]

One of his legacies, though, has been stronger than any other: Kohlberg was explicit that he was attempting to tether tightly an account of moral development and moral cognition to an account of moral epistemology and metaethics. What this meant for Kohlberg was that any adequate *explanation* of the psychological mechanisms that produce behavior identified as "moral" would be *broadly the same* as any adequate *justification* of certain principles or modes of reasoning as moral. Kohlberg (1971) provides some further clarification as to what counts as 'broadly the same':

> [W]e do hold a stronger position, claiming that while psychological theory and normative ethical theory are not reducible to each other, the two enterprises are isomorphic or *parallel*. In other words, an adequate psychological analysis of the structure of a moral judgment, and an

[17] The initial impulse to write this paper came from a discussion in Cristina Bicchieri's fall 2012 class at the University of Pennsylvania on moral psychology. Reflecting the interdisciplinary nature of the subject, the class drew students from philosophy, psychology, economics, political science, and history. During a week of studying Kohlberg and Gilligan's work, there was a lively debate in class on the continued relevance of reading Kohlberg, with (roughly) the non-philosophers arguing that Kohlberg should be ignored since *he got the facts wrong* and with the philosophers and historian responding that it is important to understand where the discipline came from, that there is a complicated relation between conceptual schemes and facts, and (in the case of one person) that there are no *facts per se*. My response was that we should read Kohlberg because, without being fully aware of it, we could inherit some of his conceptual baggage, and we might not want some of that conceptual baggage. This paper extends that point – what we as a discipline have inherited from Kohlberg is a rather odd idea about the relation between descriptive psychology and normative theorizing, that each is constrained by the other. And, I argue, we should not accept that.

> adequate normative analysis of the judgment will be made in similar terms. In the context of our work, psychological description of moral stages corresponds to the "deep structure" of systems of normative ethics. The logical relations between stages represent indifferently the structure of an adequate theory of moral judgment development, or the structure of an adequate theory as to why one system of moral judgment is better than another. Thus, we have argued for a parallelism between a theory of psychological development and a formalistic moral theory on the ground that the *formal psychological* developmental criteria of differentiation and integration, of structural equilibrium, map into the *formal moral* criteria of prescriptiveness and universality. (p. 224)

The particular details of Kohlberg's account – related to differentiation, integration, prescriptivity, and universality – need not concern us at the moment. What is important to take from the passage is that, for Kohlberg, the *logical structure* of the true descriptive account is isomorphic or parallel to the logical structure of the right normative account,[18] such that (i) each discourse contains

[18] I apply "true" to the descriptive account and "right" to the normative account in order to emphasize that Kohlberg still accepts some minimal conception of the fact/value distinction, such that there are some instances of naturalistic fallacy (of course, none from his own theory, according to Kohlberg). Kohlberg tries to bridge the gap between the 'true' and the 'right' by means of a pragmatic standard of 'adequacy.' But, "adequacy" itself is an evaluative notion and a slippery one at that. "Adequate for what and to whom?" we should ask.

the same or "similar" terms[19] and that (ii) descriptions of *types* of moral judgment directly map onto justifications of particular morality systems (and vice versa).[20]

In this paper, I will argue that Kohlberg's assumption is deeply wrong. While I accept that findings in descriptive psychology should impact our commitments to particular normative theories, I do not accept that our particular normative theories should impact our commitment to accounts of descriptive psychology, at least not in the strong way he demands.

We should not accept the latter because it is possible to develop an account of moral psychology that is *minimalist*. By "minimalist," I mean that it is possible to theorize about descriptive moral psychology without importing substantive and ultimately tendentious normative commitments. Since it is possible to do that, we are not forced to accept Kohlberg's assumption. We can

---

[19] Kohlberg in the passage quoted above on this page talks about "similar" terms but in other places (see the epigraph to this paper) talks about sameness. This terminological fuzziness places pressure on his claim that there is an "isomorphism" between psychological theory and normative ethical theory, for I take it to be sound to say that if X is similar to Y, then X and Y are different in at least one respect other than numerical identity (because if they were not different in at least one respect other than numerical identity, then they would be the *same* – speaking at the type rather than token level—and not similar). So, the similarity relation is not "edge-preserving," so to speak. Still, we need not get too hung up on this point. For more, see (Goodman 1972).

[20] "Morality system" is a philosophical term of art from Bernard Williams, first developed in the late 1970s before fuller development in the 1980-90s. For our purposes, allow "morality system" to pick out theories of morality as developed by Western philosophers and theologians. Paradigmatic examples of the morality system includes divine command theory, consequentialism (including utilitarianism), deontology (including contractualism and contractarianism), and (in its ambitious and universalistic forms) neo-Aristotelian virtue ethics. To be sure, these examples share common features that make them fall under the concept 'morality system,' but those features are orthogonal to my argument here. For more on Williams's specific notion, see (1985).

then choose between accepting the assumption on the basis of pragmatic or meta-theoretical considerations. And there are two strong meta-theoretical considerations that weight against the assumption: namely, the incoherence of morality systems in general as a result of social and cultural evolutionary history and the extraordinary "lightness" of metaethical philosophizing.

The structure of the paper will be as follows: first, I will further explicate Kohlberg's assumption – its motivations and its consequences – through a close philosophical reading of "From Is to Ought." Next, I will argue that we should not accept the assumption, for reasons outlined above. Finally, I will conclude with some considerations about the relation between realism and naturalism in moral psychology, given a rejection of the assumption. A continued theme of this paper is that we need a moral psychology, not a moralized psychology. By this I mean to give slogan to the *minimalist moral psychology* partially described above. But, if moral psychology is just to be the psychology of moralizing or certain kinds of moralizing or the psychology of entertaining particular substantive normative commitments,[21] then perhaps we would be better off without moral psychology as well.[22]

2 The Assumption: Motivations

[21] Such that "moral psychology" becomes a base kind of cognitive phenomenology, or what-it's-like to think that P (where P is some substantive moral claim, like "justice is the supreme value" or "avoiding harm is the trumping obligation" or "it's wrong to torture cats for fun").
[22] I will not argue explicitly for this claim, but it will be implicit in much of the discussion to follow.

It is important to note at the outset that Kohlberg's assumption is, in fact, an *assumption*. There are two ways this is so: he does not provide an argument for the claim, and there are *reasons* the claim *must* be an assumption – there is no framework from which we can say, *ex ante* to the deliverances of his moral psychology, that descriptive psychology and normative theory go hand-in-hand. Because of his pragmatist orientation, Kohlberg must be committed to a view where his claim can be *vindicated* by means of the *results* of the theory. To put the point metaphorically, if the proof of the pudding is in the eating, the claim "the pudding is delicious" can *only be* vindicated after tasting it (and *in virtue* of how it tastes), never demonstrated or compelled by considerations prior to the eating.[23]

But, given that the assumption is an assumption, why would anyone assume *that*? I propose that there are two reasons Kohlberg made the assumption. The first is a reason related to history of the practice of psychology, the second related to Kohlberg's substantive normative commitments.

Kohlberg saw the evolution of developmental psychology in the 20th century as moving away from behaviorist models to cognitivist models. Behaviorist models, focused on stimulus and response, provide an explanation where it is assumed that

> the process of learning truths is the same as the processes of learning lies or illusions. It explains the learning of logical operations or "truths" in

[23] On Kohlberg's acceptance of his claim as an assumption and his acceptance of pragmatic vindication, see p. 225.

> terms of the same processes as those involved in learning a social dance step (which is cognitively neutral), or those involved in "learning" a psychosis or a pattern of maze errors (which are cognitively erroneous). (1971, p. 152)

The scare quotes are illustrative. First, learning a dance is not a matter of truth at all, according to Kohlberg. Presumably, he has a conception of truth where truth is a property of propositions, specifically of propositions of the form "S knows that P."[24] A dance is not a proposition – learning the pattern of bodily movements that are constitutive of the dance such that you can perform the dance does not *represent* the world as being such-and-such way. To be sure, there is *a pattern*, and one can learn – in a straightforward sense of "learn" – the pattern. But, to learn the dance would be an instance of *knowledge-how*, not *knowledge-that*. Presumably, then, only the contents of 'knowledge-that' count as truths. Second, things that are "cognitively erroneous," like lies or illusions or ways through a maze that do not get you out of the maze, are not things that can be properly "learned." This suggests that we can only "learn" that which is, in fact, true.

By hypothesis, then, we can come to believe that, say, the present king of France is bald, but, if in point of fact there is no king of France presently, then we did not learn what we believe. Putting these two points together, we can say that,

[24] Strictly speaking, the proposition would be "that P," to which the propositional attitude "knowing" attaches. I will treat the entire expression as a proposition since I am not particular interested in nor is it relevant to my argument to discuss the metaphysics of proposition-hood.

for Kohlberg, we can only learn what is true and that the only things that can be true are propositions.[25]

But, behaviorism was unable to distinguish between the three cases, between the "cognitively true," the "cognitively erroneous," and the "cognitively neutral." Specifically, behaviorism could not pick out the "cognitive" as a distinct psychological process. But the pioneering work of Piaget on the childhood development of *concepts* like space, time, and causality demonstrates that explaining the behavior of children requires philosophical, epistemic notions (again, like space, time and causality). (Kohlberg 1971, p. 152). So, a psychology rested on behaviorist epistemology was insufficiently explanatory: it did not explain the types of things that we know to exist. We *know* that children have certain ways of getting around the world that are more successful than others. How do we explain that? –By invoking a cognitive-psychological mechanism that makes possible such developments.

Kohlberg think that since such an explanation was in the offing in one area of developmental psychology, it should be in the offing in all areas of developmental psychology. This assumption provides one basis for *the assumption*, the assumption at issue in this paper about the relation of descriptive psychology and normative theorizing.

[25] We can of course relax the standard to include belief-states that do not have fully explicit propositional form. We could have inchoate beliefs, and those could still be true, for they are truth-apt (by the hypothesis that for a mental state to be a belief state is for it to be truth-apt).

Of course, all this groundwork is to a point already assumed in the discipline: if there is *development*, then the development develops *toward something or other*.[26] But this high level of generalization is insufficient for Kohlberg. It cannot be the case that our moral cognition – or even cognition more generally – develops toward something or other. It has to develop toward *some thing*. That particular thing is going to be the *ideal*. At the highest level of generality, the ideal is the example *par excellence* of whatever thing is under consideration – it is *that thing* that best exemplifies what makes a thing of that type *a thing of that type*.[27]

In the realm of moral cognition, this means that there must be an *ideal* form of moral thinking. That is, there is a form of moral thinking that best exemplifies what makes moral thinking *moral thinking*. That "best exemplification" or ideal specifies (a) what moral cognition develops toward by (b) giving a standard against which to *measure* (c) either *alignment with* or *deviation from* the standard. At any point along the development trajectory, the next level of development *contains* the previous level and *adds* something distinct. There is need of a next level if, at a level, that level that the development is at is unable to solve for problems that the level identifies as problems. The development trajectory halts just in case there is a level of development that

[26] A tautology, to be sure, but a useful one. It provides *prima facie* evidence that further refinements move away from tautology toward the substantive. The substantive can always be denied on the grounds of empirical falsity or conceptual incoherence.

[27] This is truly at the highest level of intelligible generality. "Thing" here could pick out objects of different kinds – physical and intentional – as well as events, sequences, orders and the like. My use is catholic and ecumenical.

solves all problems of previous levels while not encountering any problems that are recognized as problems *from that level*.

But, how do we find out which form of moral thinking provides that sort of "best exemplification?" One way – a pragmatist way, Kohlberg's way – is to ask, "What is moral thinking *for*?" We try to figure out what role or function morality performs for those who have morality. Given that only humans have morality, we try to figure out what problems does morality allow humans to solve. Kohlberg (1971) has an answer: the function of moral cognition is to resolve moral conflicts with others, to give guidance in how we should act in different environments, to eliminate moral dilemmas, and to do these things in a *stable* and *consistent* way (p. 185). The requirements of stability and consistency in resolving moral conflicts give rise to demands for a formalistic metaethic with certain substantive content claims. The substantive content claims relate to the value of persons and to the supreme overriding value of justice. The particular details need not bother us here.

To summarize, there were two reasons Kohlberg makes the assumption, one related to developmental psychology as a discipline and one related to the substantive normative commitment he held. The second reason is intimately tied to the consequences of the assumption, so although I have broached the topic in this section, my extended discussion takes place in the next.

3 The Assumption: Its Consequences

It just so happens that Kohlberg's identified function has the salubrious effect that it vanquishes a range of views that Kohlberg (1971) does not accept: descriptive relativism (p. 176),[28] normative relativism (p. 180),[29] emotivism (p. 184),[30] epistemological intuitionism (p. 184),[31] motivational internalism (pp. 217-218),[32] and critical / analytic metaethics (pp. 224-225).[33] Also vanquished is the normative theory that corresponds to each level of moral thinking that is *lower* than Stage 6 theory (1971, p. 216): Stage 1's rule-and-authority obeying morality, Stage 2's rational egoism, Stage 3's commonsense morality, sentimentalism and virtue ethics, Stage 4's conventional morality and rule-and-authority maintaining morality, and Stage 5's rule-utilitarianism, social contract theories,[34] and

---

[28] Descriptive relativism is a moral metaphysic that says that morality differs from culture to culture.

[29] Normative relativism is a normative doctrine that says that we should not judge people from cultures that have different standards than our own.

[30] Emotivism is a semantic theory that says that the meaning of moral terms is captured completely by their emotive content. More specifically, emotivism holds that moral language expresses or evinces speaker attitudes, where the attitudes are taken as non-cognitive. For the classical position, see (Ayer 1952).

[31] The term "epistemological intuitionism" comes from Williams and is meant to contrast with "methodological intuitionism." Epistemological intuitionism posits that we intuit or directly apprehend moral facts by means of a special epistemic faculty. G. E. Moore's metaethic is a classic example. See (Williams 1998b) and (Moore 1996).

[32] Motivational internalism is a moral metaphysic that says that morality is, in some way to be specified, inherently motivating. In stronger forms, it denies the possibility of the amoralist – someone who says, "I understand X is right, but why should I care?" Kohlberg's theory is externalist, in this sense.

[33] Critical/analytic metaethics is a semantic and epistemological theory that says that the task of moral philosophy is to clarify the principles that are already implicit in "ordinary" moral language.

[34] Some people take Kohlberg to be developing a theory of moral cognition that is roughly contractarian. This is a mistake. He writes,

> At stage 5, the core of justice was (a) liberty or civil rights, (b) equality of opportunity, and (c) contract. These three ideas were united by respect for the freedom of others, as this freedom is embodied in civil law and civil

methodological non-relativism. The last theory standing is the morality of Stage 6 morality: deontological, or principled, intuitionism (1971, p. 212, p. 219).[35] What is remarkable is the *range* of types of views that Kohlberg's assumption plus theory licenses him to strike: moral-semantic theories, moral-epistemological theories, moral-metaphysical theories, moral-motivational theories, descriptive psychological theories, and substantive normative theories.

But what is philosophically interesting is that Kohlberg's moves to strike rival theories only work in the context of identifying standards of metaethical correctness, and the only way to identify standards of metaethical correctness is by appeal to substantive normative theory.

> rights. At stage 6, the sense of justice becomes clearly focused on the rights of humanity independent of civil society. (p. 212)

Or consider this quote: "We have been arguing that, both by stage 6 normative ethical standards and by formalist metaethical criteria, stage 6 is a more moral mode of judgment than stages 5 or 4" (p. 217). He does allow that Rawls derives Stage 6 morality from Stage 5 morality, but only insofar as the morality pertains to social-political choices (which does not exhaust the content of Stage 6 morality) (p. 226).

[35]Kohlberg invokes as representative of Stage 6 morality the kind of morality endorsed by Ross and Sidgwick. His claims about motivational externalism and the impossibility of answering the amoralist in non-moral terms call to mind another deontological intuitionist – namely, Prichard. See (Ross 2002), (Sidgwick 1981), and Prichard (1912). All in all, Kohlberg's theory amounts to a rather boilerplate recapitulation of the type of deontological moral theorizing prevalent in the early 20th century between the publication of Moore's *Principia* in 1903 and Ayer's *Language, Truth, and Logic* in 1936. Of course, Sidgwick came before this period and tried to defend consequentialism. But, he accepted a kind of principled intuitionism, which Kohlberg notes even if some contemporary moral psychologists – less philosophically able than Kohlberg – deny that Sidgwick used *intuitions* at all (p. 219). Kohlberg's philosophical mistake is to identify any reliance on principles and formalism as underlying a *deontological* position. Deontology just is the normative theory that says that an act is right iff it accords with the right moral rule. Sidgwick never accepted that.

Substantive normative morality is the "thick" stuff of the moral life: the particular and substantive commitments that we have as moral agents embodying a particular moral worldview. For example, a hedonistic maximizing utilitarian might have the particular and substantive commitment to eradicate factory farming, where the content of her normative reason consists of the idea that eliminating factory farming will move the world from one state of affairs to another and that the latter state of affairs contains more overall utility as measured by pleasure/pain indices. This utilitarian may have a *realist* metaethic, such that she considers her normative reason *formally* as a mind-independent fact and as on a par with the other sorts of facts given realistic interpretations in, say, the natural sciences. Her explanation of why her normative reason is *normative* is that *it is true* that factory farming is wrong.

Kohlberg's theory is like that, not in the sense that his is realist utilitarianism but in the sense that substantive morality and metaethical theorizing go hand-in-hand. He claims that morality is *sui generis* and formally autonomous – that is, morality is not a subset of any other domain. Because of this metaethical "fact," Kohlberg (1971) can say things like,

> The general criterion we have used in saying that a higher stage's mode of judgment is more adequate than a lower stage is that of morality itself, not of conceptions of rationality or sophistication imported from other domains. (p. 215)

or like, "We have been arguing that, both by stage 6 normative ethical standards and by formalist metaethical criteria, stage 6 is a *more moral* mode of judgment than stages 5 or 4" (p. 217, emphasis *mine*). And what of this criterion of "morality itself?" We get to it by means of the formal metaethical characterization of the features of moral thought. The features that we build into the formal characterization *constrain* the choices at the normative level. For Kohlberg (1971), the metaethic uniquely determines at least some of the substantive normative content to which our moral theory is committed:

> If our formal characterization of the functioning of mature principles is correct, it is clear that only principles of justice have an ultimate claim to being adequate universal, prescriptive principles. By definition, principles of justice are principles for deciding competing claims of individuals, for "giving each man his due." When principles, including considerations of human welfare, are reduced to guides for considering such claims, they become expressions of the single principle of justice. (pp. 219-220)

Kohlberg then goes on to discuss taking considerations of human welfare as an alternate content claim before rejecting the position for failing to satisfy the metaethical constraints of prescriptivity and universality.[36]

But, assume that Kohlberg is right that the metaethical constraints uniquely yield the principles of justice as the correct normative principles. And

[36] In essence, arguing in the *opposite* direction – against utilitarianism – from universality and prescriptivity than the philosopher most wellknown for introducing universal prescriptivism as a metaethical theory. For that other view, see (Hare 1981).

assume that the metaethical constraints uniquely pick out his descriptive psychology. What we have, if Kohlberg is to be believed, is a case of a very powerful explanation: moral psychology interlocks with moral epistemology, moral metaphysics and moral semantics, which in turn interlocks with first-order normative commitments to a contentful principle of justice. In the end of explanation, then, moral psychology is linked to substantive normative morality, via logical connection.

4 The Assumption: Its Problems

Say that you have a particular first-order normative commitment to the principles of justice as understood at the Stage 6 level of Kohlberg's theory. In fact, say that you are deeply committed to the principles. Given that you have such a commitment, it would be convenient if there were a metaethic that uniquely picked out your commitment. Even better, what if that metaethic validated one account of moral cognition over all others? Then, in virtue of establishing the metaethic by means of laying out the descriptive psychological evidence in support of that ethic, you have laid out support for your normative commitment. You get the normative commitment *for free* by means of "logical necessity."

But, of course, you get the normative commitment for free only in virtue of making the metaethic and the psychology out of the materials of the normative commitment at hand. You start with the normative commitment, you reverse engineer a metaethic that uniquely selects your normative commitment,

and you propose a psychological mechanism that is "isomorphic" or "parallel" to your metaethic. If there really is an isomorphism, then the psychological mechanism must also uniquely select your normative commitment. One way to look at the resulting omni-theory is to see it as having great explanatory breadth and depth. Another is to see it as an instance of circular reasoning: your normative commitment allows you to rule out certain metaethics, which allows you to rule out certain psychological mechanisms, which allows you to rule out certain metaethics, which allows you to rule out certain other proposed normative commitments. But, at no point in the "explanation" have you provided anything like a reason to accept the omni-theory for someone *not already* in the grip of the normative commitment.

One consequence from the assumption is that it allows the theorist to strike all sorts of rival theories, theories in the psychological, metaethical and first-order normative domains. A second consequence is that it allows for the possibility of the theorist reverse engineering an explanatory and justificatory framework for the particular normative commitments she happens to hold. A third consequence is that the assumption promotes a circular theory, although whether you find the theory virtuously or viciously so will depend on whether you accept the point of entry.

I do not mean the preceding paragraphs as an opening salvo in yet another round of the most boring topic in all of moral philosophy: who has the right definition of morality? That topic – definitions of morality and their taxonomies – is of vanishingly small importance, for people *go on* in the absence

of such definitions.[37] Rather, I suggest that something like the paragraph above offers partial explanation for Kohlberg making the assumption at issue for this paper. Of course, this is speculative psychology, and there can be no apodictic philosophical demonstration of such speculation. But, and this is the upshot, it is *explanatory* of why someone would make Kohlberg's assumption. As Williams gestures at in the epigraph to this paper, when a psychological mechanism that has a unique fit with a moral conception is proposed, we do well to wonder if the mechanism itself is part of the moral conception. If it is part of the moral conception, then why should we accept it as a matter of *psychology*? To jabber on about the *sui generis* and autonomous nature of *morality* is, at that point, to severely *miss the point*. So too would giving an architectonic of the thirteen or thirty[38] definitions of morality from the history of moral philosophy.

Skepticism about psychological mechanisms that have unique fits with moral conceptions rest upon skepticism about the moral conceptions. Another way to put the point is to say that if the mechanism implies the conception, then attacks on the conception imply attacks on the mechanism.[39] If commitment to the moral conception is not compulsory – not *obviously* right – there is a lurking problem for the descriptive theory.

5 The Minimalist Alternative

[37] Although it is interesting to observe that some of those most heavily invested in preserving the *practical* status of morality are so concerned with making sure that everyone *have the right definition* in mind. Often, arguments of this sort are unclear about the form/content and theory/practice relations.
[38] Or however many artificially selected . . .
[39] An application of *modus tollens*.

This is not to say that there is a theoretically neutral way to characterize the psychology of moral cognition. Nor is it to say that we can excise *all* of our normative commitments in developing our descriptive theories. My call is not a call for vulgar positivism or dogmatic intuitionism. Rather, I suggest instead a *minimalist moral psychology*.

Borrowing from Williams,[40] we can say that a moral psychology is *minimalist* iff it satisfies two conditions. The two conditions go together, but I will discuss them in turn.

The first condition relates to how much moral content we should place into our account of the psychology of human beings. Williams (1998b) writes:

> First, to the question 'how much should our accounts of distinctively moral activity add to our accounts of other human activity?' it replies 'as little as possible', and the more that some moral understanding of human beings seems to call on materials that specially serve the purposes of morality – certain conceptions of the will, for instance – the more reason we have to ask whether they may not be a more illuminating account that rests only on conceptions that we use anyway elsewhere. (p. 68)

Williams in this passage is talking about conceptions of the will like Kant's, conceptions where there is always a double-action.[41] But, the particular details

[40] Who in turn borrowed from Nietzsche.
[41] The point: for each action, that account of willing adds another – namely, the action of willing! The action of willing is marshaled as an explanation of action, but it cannot *really* serve as an explanation of action (any more than Unmoved Mover arguments "explain" the sequence of events in the natural order). The

need not concern us. What is relevant is that we have an account of human cognition. Of this account of cognition, we should import as little specifically moral material as possible in constructing it. This, too, should hold for our accounts of moral cognition. We try to provide accounts of moral cognition: how much should our *descriptive* account of moral cognition *import* from our *normative* commitments as agents involved in interacting with others as *moral agents*?

Kohlberg's answer: "Quite a bit." We import a conception of justice linked closely to a very special notion of human beings as free and equal moral persons with inviolable dignity. We import a conception of moral judgment with strong demands related to universalizability, prescriptivity, and so on. We import a conception of the relation of the two where the formal features of moral judgment uniquely yield the substantive content that is the first-order commitment to justice.[42]

Of course, Kohlberg (1971) claims that his metaethical conception of the later stages as more moral than the earlier stages does not amount to a normative ethical principle (p. 217). And he claims that his Stage 6 principles do not *directly* require any rule of action or theory of the good (p. 217). But each claim is simply not true and simply not true on his own account. The metaethical conception does amount to a normative ethical principle: since anyone at a stage can understand the reasoning of any stage lower and since occupying a stage does

---

question arises, "What explains the willing?" and there is no good answer for *that* in the offing. Hence, Kant invokes transcendental psychology and noumenal purposes, explaining the difficult by means of the baffling. See (Kant 1998), especially the third section.

[42] As understood in its Stage 6 interpretation.

not directly determine[43] that you will think only in terms of that stage, it is possible, at any stage above the first, for you to *choose* your approach the structure with which you are trying to achieve equilibrium. Say I am at Stage 3. I can choose between deciding to do something on the basis of virtue ethical considerations or deciding to do something on the basis of rational egoistic considerations (i.e., Stage 2). Kohlberg's theory says that I should deliberate from Stage 3. It is normative: it "authoritatively" tell me which path to pick among options and "resolves" the "moral conflict" of choosing between deliberative stages.

Consider now the Stage 6 principles of justice: they say that it is best to regard people as free and equal moral persons with inviolable dignity. It defies my comprehension that this does not amount to a rule of action. If I can choose between an action in accordance with the principle and one not, I should always chooses to act in accordance with the principle, *according to the principle*. If there is a preponderance of principles, I should choose the action that satisfies more of the principles than any other available course of action. I see no other way to read the claim.

This last paragraphs shows how much the assumption assumes. It assumes quite a bit of distinctively moral content – about the nature of obligations, about the moral nature of human beings, and about permissible

[43] That is, with causal necessity.

action. This is to say that contrary to Kohlberg's claims, his theory does have something to say about the nature of the good and the nature of approbation.[44]

I will give a first approximation as to what counts as importing as little distinctively moral content in a bit. But, before I do, I must lay out the second aspect of the minimalist position, for the second aspect goes some way toward that first approximation. Williams (1998a) continues:

> This demand for moral psychological minimalism is not, however, just an application of an Occamist desire for economy, and this is the second aspect of the Nietzschean general attitude. Without some guiding sense of what materials we should use in giving our economical explanations, such an attitude will simply fall back into the difficulties we have already met. Nietzsche's approach is to identify an excess of moral content in psychology by appealing first to what an experienced, honest, subtle, and unoptimistic interpreter might make of human behavior elsewhere. (p. 68)

[44] Namely, that it is *good* to do the *right* thing by treating other humans as free and equal moral persons and that punishment may be required, as a matter of fact given a limited set of available actions, for distinctively moral reasons not related to concerns of social utility (where social utility is read in a consequentialist, read: non-principled, way). That a deontological theory of the right has plenty to say about the good did not strike other early 20th century deontological or principled intuitionists as wrong. See (Ross 2002). The whole point of Ross's theory was to invert Moore's analysis. (Moore 1996) says that 'good' is a *sui generis* concept but that 'right' is *analyzable* in terms of being productive of the good. For Ross, 'right' is the *sui generis* concept, and 'good' is analyzable in terms of being productive of the right. This rather obvious fact is still accepted as rather obvious in the way that we introduce undergraduates to consequentialism and deontology: a common gloss is that consequentialism defines the right through the good and that deontology defines the good through the right.

The kind of "economical explanations" at issue here are explanations of the *naturalized* sort. Here I endorse Brian Leiter's (2002) reading of Nietzsche as primarily a soft methodological naturalist.[45] Soft methodological naturalism holds that philosophical enquiries should be continuous with the methods of successful natural and social sciences, including "styles of explanation and understanding employed in the sciences" (Leiter 2002, p. 4). Nietzsche is a substantive naturalist with respect to ruling out all forms of supernaturalistic – theistic or deistic – explanation. I take Williams's position to be the same, and I too subscribe to soft methodological naturalism combined with substantive naturalism about theological talk.

The "difficulties we have already met" refers to the difficulties in general with provided a *naturalized moral psychology*. These are the particular difficulties that I have been claiming beset Kohlberg's theory. Williams (1998a) again:

> If a 'naturalistic' moral psychology has to characterize moral activity in a vocabulary that can be equally applied to every other part of nature, then it is committed to a physicalistic reduction that is clearly hopeless. If it is to describe moral activity in terms that can be applied to something else, but not everything else, we have not much idea what those terms may be, or how 'special' moral activity is allowed to be, consonantly with naturalism. If we are allowed to describe moral activity in whatever terms

[45] See especially chapter 1, "Introduction: Nietzsche, naturalist or postmodernist," pp. 1-30.

> moral activity may seem to invite, naturalism excludes nothing, and we are back at the beginning. (p. 67)

We have three options for naturalizing moral cognition: in terms of a general purpose discourse that applies to every other part of nature, in terms of whatever discourse moral cognition seems to invite, or in terms of some in-between discourse – a discourse where some things apply but not all. I agree with Williams that the first option is hopeless: the only candidate option is theoretical physics, and it would both be a fool's errand and a serious misunderstanding of moral thought to try to characterize it in terms of fundamental particles (or their fields) worked upon by forces understood as laws of nature. It is the wrong level of explanation altogether. So, that only leaves two options: allow in whatever discourse morality requires or seems to require, or describe morality in a discourse that is outside of morality yet the universe of which is "suitably restricted."

Kohlberg opts for the first of the remaining two options: in order to explain moral cognition, he will invoke whatever terms moral cognition seem to him to invite. This includes the terms in his formalistic metaethics and the determinate and substantive content of Stage 6 morality. It also includes the determinate and substantive content of each stage prior to Stage 6. That is to say, in explaining what makes the lower stages *stages*, given the assumption, Kohlberg calls on the determinate and substantive content of all possible morality systems and folk moral worldviews. Strictly speaking, rational egoism, virtue ethics, conventionalism, legalism, social contract theories, and

utilitarianism are all present in the theory, if only to be subordinated under the master notions of justice and deontological intuitionism. He excludes *nothing*: protests that pre-conventional moralities (Stages 1 and 2) are not proper moralities (because not properly principled) are themselves unprincipled, given that each higher stage must recognize the previous stage as something *from which it sprang*.

The last remaining option is the only real option for naturalistic moral psychology, hence for moral psychology. But, the last option directly cuts against the assumption. If we do not know how the restrict our universe of discourse, then how can we know, either in advance or as a result of "investigation," which terms are the genuinely referring terms that carry sense? Does moral cognition have to be universal and prescriptive? If so, is there a basis for *that* besides antipathy for relativism?

I should quickly note that Kohlberg's strategy with regard to relativism is to show that all versions rest on different "logical fallacies."[46] That is an insufficient strategy. Relativism may well rest on a mistake and be conceptually

[46] Another unfortunate habit that Kohlberg picked up from early 20th century metaethicists is the habit of labeling all opponent views as resting on a fallacy or a mistake. Besides expressing a generally negative outlook about the cognitive abilities of the interlocutor, such moves also involve some narcissistic preening. See the end of (Kohlber 1971) where he is discussing truths passed down from Socrates, truths that psychologists have not accepted. Kohlberg's explanation: "Is it so surprising that psychologists have never understood Socrates? It is hard to understand if you are not stage 6" (p. 232). That statement has pretty sour implicature: either Kohlberg is not stage 6 but is smart enough to overcome the deficiency (unlike all his opponents), and he is stage 6 and so is both smarter *and* better than his opponents.

incoherent.[47] That does not show that universalism is conceptually coherent. You need additional premises: relativism and universalism are the only options, they are mutually exclusive, they are opposite such that the falsity of one implies the truth of the other, and so on. And you need to exhaust all possible logical types of relativism and universalism. Needless to say, Kohlberg did not accomplish that daunting task. And, there may be *in principle* reasons why he could not. After all, if, as Kohlberg claims, moralities are constructions putting responses in equilibrium with structures, then, as environments and material conditions continue to change, what counts as "equilibrium" is also subject to that change. This is another way of saying that Kohlberg's identification of the *function* of morality is merely a reflection of his own personal predilection.[48]

The important point is that we are not compelled to accept Kohlberg's analysis. There is a minimalist option of the table. The minimalist option need not necessarily be relativist: it is possible to develop a universalist minimalism.

---

[47] What I call "vulgar relativism" is surely incoherent: P1. Morality varies across cultures. P2. There is no overarching standard by which to judge among other cultures. So, C., Don't judge and be tolerant!

[48] This is yet another reason to avoid the most boring question in moral philosophy. We can allow that "morality" means this or that or the other while still saying that the function of morality is such-and-such. To bring in an illustration from perception, we can say that "X looks red" *means* "X has the micro-physical structure corresponding to red" or "X seems red to person P in circumstance C" or "Red is the quality of my sense-datum in relation to perceiving thing X." Regardless, we could still identify the function of X looking red by means of the causal role looking red plays in behavior (say, in identifying pomegranates). But, where there are multiple possible functions in play and where the function is given a self-referential role, then assigning non-reductive content to the function becomes problematic. Example: the function of morality could be to allow the weak to keep the strong in check, or the function of morality could be to allow us to solve moral conflicts. On the basis of *what* do we identify the function? On the basis of what do we say that there is a "*the* function?"

The universalist minimalism imports very little into the psychology: it would not import, for example, universal prescriptivism or full role reversal. It would *explain* particular moral judgments by means of other ready-to-hand tools that do not have distinctive moral content: here I am thinking of theories of social norms, decision theory, theories of rationality, theories of politics and so on.[49]

6 Meta-Theoretical Considerations in Favor of Minimalism

So, we have to choose between the Kohlbergian assumption and the minimalist paradigm. I argue that there are two meta-theoretical considerations that should have some weight in pushing us toward minimalism: first, morality systems in general are incoherent, and second, metaethical theorizing, once already in the naturalistic purview, is unlikely to have any effect on first-order cognition.

First, morality systems in general are incoherent. From a naturalistic perspective, this is unsurprising. If morality systems are the products of biological and cultural evolution, reflecting a line of constant adjustments from different pressures, then we ought to expect that morality cannot be reflectively coherent. Outside of the morality systems themselves, the stuff of ethical life – our thoughts and practices related to getting along with along humans with killing them – is itself not the kind of thing that can hang neatly together. If we assume, as seems entirely reasonable,[50] that there are ineliminable moral dilemmas in the actual world and if we combine that assumption with the

[49] On social norms, see (Bicchieri 2005).
[50] See (Marcus 1980).

conception of morality as dilemma-solver in the actual world,[51] then morality is more likely than not incoherent. It probably cannot perform what it claims is its function to perform. We eliminate the incoherence by abandoning one of the claims, and that move will reflect antecedent normative commitments that we have about the nature of obligation as such.

Besides coherence, there remains the fact that metaethical theorizing, like many kinds of philosophical theorizing, is extremely "light." What I mean by this is that these second-order theorizings rarely have first-order effects. Think of Hume and his account of causality. Hume's second-order skepticism did not prevent him from judging accurately the trajectory of billiard balls at the bar. Would second-order optimism have made much of a difference? Likewise, Kohlberg himself argues that, even though there is a difference between different philosophical conceptions of 'morality,' there are no fundamental differences between those philosophical conceptions when those conceptions are measured as against psychological conceptions. Now, I find this implausible, for a naturalized position would be a philosophical conception but would look an awful lot like a psychological conception. Still, given what we know about the relations in general between first and second-order theories, we should be highly skeptical of any claim giving priority to a second-order theory to uniquely determine first-order content about which the second-order theory theorizes. Humean skepticism with regard to causality does not issue in first-order

[51] One of Kohlberg's assumptions.

skepticism that if I drop this cup, it will fall to the ground. What special reason is there for assuming the moral case is different than that?[52]

I conclude that there has not been good reason to accept the assumption. There is an alternative on the table – minimalism. Meta-theoretical considerations favor minimalism over the assumption. This considers are not positivist applications of a principle of parsimony; rather, they emerge from a meditation on the substance of the moral life as lived. Theory cannot assume away the world. Nor can it make unintelligible practice.

7 Conclusion: Realism and Naturalism in Moral Psychology

In my paper, I showed that Kohlberg had a strong assumption about an isomorphism or parallelism between descriptive accounts of the development of moral cognition and normative accounts of (a) the right formalist metaethic combined with (b) certain substantive normative claims about the nature of justice and persons. And I argued against it on the basis of an existing alternative and metatheoretical considerations. But, moral psychology as a field has moved far past Kohlberg, and today the field is one generating great enthusiasm and interests across academic disciplines.

---

[52] Again, this is with the caveat that we are already working within a paradigm of naturalistic moral psychology. If you have subscribe to a *super*naturalistic metaethic, I am more than willing to admit that giving up *that* metaethic would have profound and pervasive effects on the contents of your first-order beliefs. But I believe that only on the basis of lots of empirical evidence – social-scientific, personal-anecdotal, and cultural-historical. But a lot of the case from certain kinds of constructivists and expressivists on this matter about how metaethical positions are kinds of normative positions just emphasizes the fact of the normative horse pushing all the carts.

There are at least two kinds of moral psychology: first, a psychology that takes as its subject matter those behaviors identified as "moral," and, second, a philosophy that takes as its subject matter those theories identified as "moral." In each kind of moral psychology, then, a practitioner must have some antecedent conception of "moral" that she brings to bear on her subject matter.[53]

However, when philosophers talk about 'moral psychology,' they mean to restrict discussion to something rather particular. They often mean to talk about moral psychology as a way of *ruling out* certain normative ethical theories. The argument frequently goes like this: normative ethical theories presuppose a picture of human nature, and if particular picture of human nature presupposed by a particular normative ethical theory is false, that gives us reason to reject the normative ethical theory.

This is fine, as far as it goes. But, it does not go very far. This is related to a point that I tried to make at the end of the paper: we do not know how much to include into our universe of moral discourse. And, naturalism, *per se*, seems to be of no help in helping us get closer to an answer.

This does mean that we must make a value judgment with respect how we plan on partitioning the world. And it is here that I reassert the value of realism over naturalism. "Realism" here is not about a doctrine of mind-

[53] Some philosophers – namely Kant but also his followers, direct or indirect – took this rather plain fact as a great triumph, an indication of the autonomy of normative ethical theory from the messy realities of what used to be called "philosophical anthropology." I never understood how that inference was supposed to work, especially given Kant's embrace of "ought implies can." See (Kant 1998).

independence.[54] Recall this quote from Williams (1998a): "Nietzsche's approach is to identify an excess of moral content in psychology by appealing first to what an experienced, honest, subtle, and unoptimistic interpreter might make of human behavior elsewhere" (p. 67). The "realism" under discussion refers to the type of attitude we take our interpreter to have. She is "realistic" with respect to the true sources of human behavior, but that does not mean that she accepts nihilism or amoralism or whatever other bogeyman dreamt up by the moral philosopher. A way of being realistic is by adopting a hermeneutics of suspicion. In was by adopting this realistic method that I was able to make the conjecture that at least part of what motivated Kohlberg to accept the assumption was a particular commitment to a particular form of morality with particular content claims, related to justice and persons.

But, as Williams (1998a) is right to point out, the method works only if you are not suspicious of *everything* (pp. 68-69). In a way, it is like skepticism or projectivism. Skepticism only works if there is at least one thing you are not skeptical about, from which you can launch your skeptical doubts, projectivism only work if there is some thing *upon which* the projection *projects*.[55] These points indicate a problem with taking skepticism or expressivism as *global* attitudes, commitments, or methods. Likewise for suspicion.

Does this mean, then, that there must be at least one substantively normative commitment that we must make if we are to theorize about moral

[54] However finessed.
[55] For local moral expressivists, the "natural world" is that which is projected upon. See (Blackburn 1993).

cognition *at all*? I do not think this follows. If we read "substantively normative" as indicating the type of commitment that Kohlberg had to his principles of justice, we need not be committed to any such thing. His commitment was not defeasible, for him, as it must be for the realist.[56]

[56] I thank Cristina Bicchieri, Molly Sinderbrand, Kyle Adams for discussions related to earlier forms of this paper. All faults remain my own.

Bibliography

Ayer, A. J. (1952). *Language, Truth and Logic*. Dover. (reprint of 1946 second edition).

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Blackburn, S. (1993). *Essays in Quasi-Realism*. Oxford University Press.

Goodman N. (1972) "Seven strictures on similarity." *Problems and projects*. Indianapolis/New York: Bobbs-Merrill.

Hare, R. M. (1981). *Moral thinking: Its levels, method and point*. Oxford University Press.

Kohlberg, L. (1971). From is to out: How to commit the naturalistic fallacy and get away with it in the study of moral development. *Cognitive development and epistemology*. Academic Press.

Kant, I. (1998). *Groundwork of the Metaphysics of Morals*. 1785.

Leiter, B. (2002). *The Routledge Philosophy Guidebook to Nietzsche on Morality*. Routledge.

Marcus, R. B. (1980). Moral dilemmas and consistency. *The Journal of Philosophy*, *77*(3), 121-136.

Moore, G. E. (1996) *Principia Ethica, Revised Edition*. Thomas Baldwin, ed. Cambridge: Cambridge University Press, 1903.

Prichard, H. A. (1912). Does moral philosophy rest on a mistake?. *Mind*, *21*(81), 21-37.

Ross, W. D. (2002). *The right and the good*. Oxford University Press.

Sidgwick, H. (1981). *The methods of ethics*. Hackett Publishing.

Williams, B. (1985) *Ethics and the Limits of Philosophy*. Harvard University Press.

Williams, B. (1995a). "Nietzsche's Minimalist Moral Psychology" in *Making Sense of Humanity and Other Philosophical Papers 1982-1993*. Cambridge University Press, 182-191.

Williams, B. (1995b) "What does intuitionism imply?" in *Making Sense of Humanity and Other Philosophical Papers 1982-1993*. Cambridge University Press.

LOCATING THE NORMATIVE INSIGNIFICANCE OF NEUROSCIENCE

1 Introduction

In this paper, my broad aim is to identify some problems with the moral psychology by focusing in particular on the debate between Selim Berker and Joshua Greene. Greene has steadily produced work that draws upon the resources of both neuroscience and social psychology to give an explanation of moral psychology[57] – and, according to him, an explanation of moral philosophy itself.[58] Berker, on the other hand, is a humanist critic of Greene's approach (and, in the end, of any approach that attempts to draw upon empirical facts to reveal normative truths, insights, or predictions).[59]

The plan for the paper is as follows: first, I will present Berker's Dilemma for any account that attempts to use neuroscience (and neuroscientific evidence) to draw normative conclusions. I argue that Berker mischaracterizes Greene's position or assumes fairly strong normative commitments disallowed by the

[57] See (Greene, Sommerville, Nystrom, Darley, & Cohen 2001), (Greene 2003), (Greene & Haidt 2002), and (Greene 2007).

[58] Consider this passage (that many humanists find deeply confused) from (Greene 2013):

> At some point, it dawns on you: Morality is not what generations of philosophers and theologians have thought it to be. Morality is not a set of freestanding abstract truths that we can somehow access with our limited human minds. Moral psychology is not something that occasionally intrudes into the abstract realm of moral philosophy. Moral philosophy is a manifestation of moral psychology. Moral philosophies are, once again, just the intellectual tips of much bigger and deeper psychological and biological icebergs. Once you've understood this, your whole view of morality changes. Figure and ground reverse, and you see competing moral philosophies not just as points in an abstract philosophical space but as the predictable products of our dual-process brains. (p. 329)

[59] See (Berker 2009) and (Berker 2014).

principles of Minimal Moral Psychology. In particular, Greene is not committed to many of the invalid inferences that Berker addresses, and Greene is not trying to draw normative conclusions merely from descriptive neuroscientific premises. However, I think that Berker's Dilemma does stick to Greene inasmuch as he attempts to use neuroscientific evidence to support psychological theory that in turn is meant to support classical utilitarianism as the correct decision procedure for resolving intrapersonal and interpersonal normative disagreement.[60] When we examine Greene's arguments, we see that it is in the move to support a grand metamorality that Greene falls prey to Berker's Dilemma. I conclude by reflecting on how my diagnosis of the problems with both Berker and Greene should be, in principle, acceptable to each.

2 Summary of Greene's Original Work

A large part of Greene's earlier research, starting in 2001, attempts to bring neuroscientific evidence to bear on the Trolley Problem. For the sake of simplicity, let us say that the "Trolley Problem" is the problem of resolving the apparent conflict between judgments concerning the "Switch" and "Footbridge" thought experiments:

> **SWITCH**: "You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman. If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman. Is it

[60] That is, Berker's Dilemma sticks to the main argument of (Greene 2013) but not to earlier arguments about the harm domain.

appropriate for you to hit the switch in order to avoid the deaths of the five workmen?"

FOOTBRIDGE: "A runaway trolley is heading down the tracks toward five workmen who will be killed if the trolley proceeds on its present course. You are on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to you on this footbridge is a stranger who happens to be very large. The only way to save the lives of the five workmen is to push this stranger off the bridge and onto the tracks below where his large body will stop the trolley. The stranger will die if you do this, but the five workmen will be saved. Is it appropriate for you to push the stranger onto the tracks in order to save the five workmen?"

Most respondents say "Yes" to Switch but "No" to Footbridge. The interesting question, then, is what makes the two cases distinct?

As Berker (2009) correctly notes, the traditional response in much of the philosophical literature prior to Greene was to assume that the judgments about the cases were correct (pp. 297-298).[61] That is, assume that "Yes" to Switch but "No" to Footbridge are the correct normative answers. On such an assumption, the task of the philosopher, then, was to locate some normative principle or other that could (1) account for the difference in judgments about the cases but (2) be robust or anti-fragile enough such that the principle could survive other cases

[61] This is, I think, a fair interpretation of much of the trolley case literature, running from (Foot 1967) through (Thomson 1976), (Thomson 1985), (Kamm 1989) and others. If you are an act-utilitarian, then you should be committed to the view that answers to Switch and Footbridge should be the same. But, so long as you are not an act-utilitarian such as Unger, then you have theoretical wiggle room to come up with some other sort of account as to why judgments about the cases do not violate a norm of consistency.

without generation of counterintuitive result.[62] So, for example, the Doctrine of Double Effect satisfies (1) but fails to satisfy (2), given that many moral philosophers feel that it generates counterintuitive results in other trolley case variants.[63]

It's arguable that, to date, there has been no normative theoretical explanation put forward that satisfies (1) and (2). If that is the case, then there are two other options available: a kind of intuitionist particularism that denies the need to satisfy (2) or a positivistic descriptive accounts that seeks to identify the psychological factors – whatever they may be – that underlie the difference in judgments.

Greene's work on neuroethics[64] takes the latter option. First, he starts with a pair of distinctions. First, distinguish two classes of judgment: "characteristically consequentialist" and "characteristically deontological" (Greene 2007). Judgments are "characteristically consequentialist" if they are supported by a consequentialist criterion of right action, and they are "characteristically deontological" if they are in line with a judgment that

[62] This is just an application of the method of cases or conceptual analysis. Assume the intuitions or judgments about the original cases are correct, find a principle that captures a morally relevant difference between the cases, test the principle against other cases, and rotate cases until you refine the principle such that it covers all cases or until you run out of cases.
[63] Such as, for example, the Loop Variant. See (Thomson 1985).
[64] "Neuroethics" is ambiguous between the neuroscience of ethics and the ethics of neuroscience. Whenever I use "neuroethics" in this paper, I am always referring to the neuroscience of ethics. See (Farah 2010).

separates deontological from consequentialist judgments.[65] Under this loose schema, "Yes" to Switch is characteristically consequentialist, and "No" to Footbridge is characteristically deontological. Second, distinguish between two kinds of psychological process. There are "emotional processes" that involve behaviorally-valenced information processing that produces automatic effects, and there are "cognitive processes" that involve inherently neutral representation that do not produce automatic behavioral effects (Greene 2007).

If you combine these distinctions in the relevant way, then you get Greene's hypothesis that characteristically consequentialist judgments are driven by cognitive processes while characteristically deontological judgments are driven by emotional processes. In order to test the hypothesis neuroscientifically, Greene takes two steps: first, identify areas of the brain that other neuroscientific research implicates as necessary for emotional and cognitive processing. For simplicity, let's say that previous neuroscientific research has implicated the ventromedial prefrontal cortex and amygdala as necessary for emotional processing and the dorsolateral prefrontal cortex as necessary for cognitive processing. Then, one can test, using available neuroscientific methods such as fMRI, which brain areas preferentially respond to characteristically consequentialist and deontological judgments.

---

[65] There are an obvious number of problems with this way of marking the classes. One is marked by reference to a criterion of right action, the other by similarity or resemblance. The "characteristically" does a lot of work here as well, for there are disputes about whether the sorts of judgments Greene is interested in are those that align with judgments rendered from the theoretical perspective. But I will not delve deeper into these concerns here. For criticism of Greene on this point, see (Kahane & Shackel 2010).

A final note, which Berker also picks up on, is that fMRI is statistically noisy, so you need to test against a large number of cases that have relevant similarity to Switch and Footbridge. That requires that you hypothesize which property may make a difference between the two cases. Greene's original hypothesis was that the difference maker was "personalness." That is, Footbridge is a "personal" case, while Switch is "impersonal." Alternatively, Footbridge satisfies but Switch does not the "ME HURT YOU" criterion:

> The "hurt" criterion [= (a)] picks out the most primitive kinds of harmful violations (e.g., assault rather than insider trading) while the "you" criterion [= (b)] ensures that the victim be vividly represented as an individual. Finally, the "me" criterion [= (c)] captures a notion of "agency," requiring that the action spring in a direct way from the agent's will, that it be "authored" rather than merely "edited" by the agent. (Greene, Sommerville, Nystrom, Darley, & Cohen 2001)

With this criterion, one can generate enough cases of meant to be relevantly similar to Switch and Footbridge and then use neuroscientific methods of investigation to see if characteristically consequentialist and deontological judgments track the distinction and then whether emotional and cognitive processing corresponds to the judgments.

Greene found that personal dilemmas tended to generate characteristically deontological judgments and activate emotional processing, while impersonal dilemmas tended to generate characteristically consequentialist

judgments and activate cognitive processing. Response times for emotionally "incongruent" judgments to personal dilemmas (that is, "Yes" to Footbridge) took longer (on average two second more) than "congruent" answers ("No to Footbridge). As Berker (2009) summarizes Greene's findings: "All told, Greene et al.'s empirical results present an impressive case for their dual-process hypothesis" (p. 305).

To summarize: Greene analyzed the Trolley Problem descriptively and used available neuroscientific techniques to try to isolate a property that would allow for a descriptive account of the Trolley Problem. He found some neuroscientific evidence that characteristically consequentialist judgments tend to correlate with cognitive processing and that characteristically deontological judgments tend to correlate with emotional processing. The descriptive property that figured as the difference maker was "personalness," where this is spelled out in terms of the ME HURT YOU criterion. From this finding, Greene has argued for a dual-process theory of moral cognition, according to which there are two systems of moral judgments. One system is quick, automatic, emotional and behaviorally valenced, while the other is slow, deliberate, cognitive and representationally neutral.

3 Berker's Dilemma

We have just seen that Berker admits that Greene's finds count as evidence for the dual-process hypothesis and that the case is impressive. Although Berker does have some quibbles with some aspects of the empirical

methodology that Greene used, his primary complaint centers on the normative implications that Greene (and others) attempt to draw from Greene's work.

Famously, Greene is not just interested in neuroethics or in showing that neuroscientific evidence supports the dual-process theory. He is also interested in showing that neuroethics is normatively significant. That is, he is interested, at least partly, in showing that we have reason to discount or reject characteristically deontological judgments but not characteristically consequentialist ones.

Berker correctly notes that this would have widespread implications for normative theorizing, the most obvious of which is that most normative theorizing ought to be abandoned as it would be a *post hoc* rationalization of existing emotional biases. And Greene does push this line of argument against deontological rationalists in some of his writings.[66]

However, it is not entirely clear how exactly Greene's argument is supposed to work or even what the argument is. Attempting to pin Greene down, Berker (2009) delivers the following dilemma:

> **BERKER'S DILEMMA**: [E]ither attempts to derive normative implications from these neuroscientific results rely on a shoddy inference, or they appeal to substantive normative intuitions (usually about what sorts of features are or are not morally relevant) that render the neuroscientific results irrelevant to the overall argument. (p. 294)

[66] See (Greene 2007) especially.

The first thing to notice is that the dilemma is destructive. Either horn is fatal to Greene's larger normative project. Remember that the project is, in part, reliant upon a psychological debunking argument. In short, the relevant psychological data debunks the class of characteristically deontological judgments but leaves intact the class of characteristically consequentialist judgments.[67] Since these two classes are supposed to be exhaustive[68] of moral judgments and since we must make some moral judgment or other in dilemma cases, we have reason to rely on characteristically consequentialist judgments. However, the first horn says that this conclusion is invalidly drawn. It could be correct but not on the basis of any of the neuroscientific results. On the other horn, the neuroscientific results are wholly irrelevant because of a substantive normative intuition about which features are morally relevant. The moral intuition is doing all the work, and the neuroscience is idle accoutrement. Either way, moral philosophers need not be worried by Greene.

I have discussed Greene's work with many members of the humanist tradition, and most of these people have been unimpressed with his work. Some feel that it does not live up to the argumentative standards of contemporary Anglo-American philosophy, particularly the sort of argumentative standards that are present in analytic metaethics. But most feel that Berker effectively demonstrated that neuroscience *per se* lacks normative significance or normative

[67] And because the data for the psychological debunking story is part neuroscientific and looking for a natural kind, this kind of debunking strategy is different than a cultural debunking one. I don't have the space to further expand this point here.

[68] This is, of course, a very controversial point. Personally, I do not believe that the two classes exhaust the domain. But I grant this for the sake of argument.

upshot. But, with certain exceptions,[69] many of these humanists have been unwilling to tackle Greene's arguments head-on. Rather, these humanists have just taken as *obvious* that Greene's arguments are bad (or obviously demonstrated as bad). I will argue that this is a mistake. Humanists ought to take Greene seriously and engage with the substance of his views, rather than rely upon second-hand judgments and antecedent normative commitments as excuse to dismiss his research program.

4 Critique of Berker

With that in mind, I will turn now to my criticism of Berker. If I can show that Berker either mischaracterizes Greene's position or relies upon strong normative commitments, then I can show that we have reason to reject Berker's location of the normative insignificant of neuroscience.

4.1 First Horn Explained

Let's start with the first horn. There, Berker says that attempts to derive normative implications from the neuroscience rely on shoddy (read: invalid) inference. My strategy here is to quickly survey Berker's proposed invalid arguments and to argue that either Greene never proposed them in the first place or that Berker is reading Greene in an aggressively uncharitable manner.

[69] Along with Berker, see (Kleingeld 2014), (Wielenberg 2014), and (Lott 2016). None pursue the line of argumentation I provide in this paper of addressing in particular the problematic descriptive component of Greene's program – the characterization of System 2, which I discuss later – to provide the humanistic critique.

Here is the first example of shoddy inference Berker (2009, p. 316) proposes:

> **REASON GOOD, EMOTION BAD**
>
> P. Deontological intuitions are driven by emotions, whereas consequentialist intuitions involve abstract reasoning.
>
> C. So, deontological intuitions, unlike consequentialist intuitions, do not have any genuine normative force.

The problem, obviously, is that the conclusion does not follow from the premise. You need a bridging premise – perhaps: "all and only intuitions involving abstract reasoning but not driven by emotions have normative force." But, further, this bridging premise could be secured only on the basis of arguing against a long tradition in ethics that says that emotions disclose normative truths.[70] So the argument is invalid, and even if it weren't, it would be tendentious and unsupported.

The second example of shoddy inference:

> **ARGUMENT FROM HEURISTICS**
>
> P1. Deontological intuitions are driven by emotions, whereas consequentialist intuitions involve abstract reasoning.
>
> P2. In other domains, emotional processes tend to involve fast and frugal (and hence unreliable) heuristics.
>
> C1. So, in the moral domain, the emotional processes that drive deontological intuitions involve fast and frugal (and hence unreliable) heuristics.
>
> C2. So, deontological intuitions, unlike consequentialist intuitions, are unreliable.

[70] See (Berker 2009, p. 316).

There are two shoddy inferences here. First, identifying something as a heuristic presupposes we can tell the difference between right and wrong answer and how to reliably get to them. This is a problem with the validity of inferring C1 from P1 and P2. It could be the case that emotional processes in other domains are unreliable but not in the moral-harm domain. Second, consequentialist judgments also likely rely on heuristics, given that we are boundedly rational agents.[71] This is the inference from C1 to C2. These inferential problems are to the side of the argument that some heuristics are highly (perhaps perfectly) reliable. So the argument is invalid, and even if it weren't, it would be tendentious and unsupported.

Here is the third and final example of shoddy inference:

EVOLUTIONARY HISTORY

P. Our emotion-driven deontological intuitions are evolutionary by-products that were adapted to handle an environment we no longer find ourselves in.

C. So, deontological intuitions, unlike consequentialist intuitions, do not have any genuine normative force.

This is perfectly parallel to the first argument. C obviously does not follow from P without a further bridging principle linking environmental adaptation to normative force. And this is aside from the fact that mathematical and scientific judgments have normative force but evolutionary history. Greene could try to

[71] That is, our brains do not compute all actual and expected consequences of actions, and we have limited memory, computational powers, and so on.

save the account through an appeal to tracking, but then neuroscience drops out of the picture. So the argument is invalid, and even if it weren't, it would be tendentious and unsupported, or neuroscience would be wholly irrelevant.

4.2 Reply to First Horn

Notice that in all of the arguments, either they are invalid or tendentious while unsupported. That is, all of these arguments are constructed in such a way that they are obviously invalid. On the assumption that we should not attribute obviously invalid arguments to our interlocutors because of a demand of the Principle of Charity, we should not attribute the charge of shoddy inference to Greene.[72]

It is true that sometimes philosophers put forward invalid arguments. And it is true that there is a phenomenon of the Principle of Charity going on holiday. I mean that there are circumstances in which, in the move to be charitable toward an interlocutor, we completely misconstrue their arguments by cleaning them up for him or her.

However, I claim that the Principle of Charity demands that if we want to attribute invalid arguments to our interlocutors, we better have airtight textual evidence. I think that this is a claim that all philosophers have reason to accept. It allows us to err on the side of charity while at the same time being able to call a spade a "spade."

[72] See (Davidson 1984).

So my primary argument that Greene is not making these invalid arguments is that, if he were, surely a philosopher as careful and thoughtful as Berker would've cited instances of the invalid arguments. Berker does not cite any instances, so Greene is not making these invalid arguments.

In the end, I think that Berker agrees with this diagnosis that Greene does not land on the first horn. Berker (2009) writes:

> Before turning to Greene's and Singer's central argument against the probative force of deontological intuitions, though, I want to briefly discuss three bad arguments for that conclusion. On a **charitable interpretation of Greene and Singer, these are arguments that they don't actually make** but which it is extremely tempting to see them as making; on an **uncharitable interpretation of Greene and Singer, these are bad arguments that they sloppily mix in with their main argument**. My guess is that **the truth lies somewhere in between**: although Greene's and Singer's primary and most promising line of argumentation does not rely on these three arguments, **I think they occasionally give their main argument more rhetorical force by invoking versions of these arguments. So it is worth showing just how unconvincing these three arguments are** before we consider Singer's and Greene's main reason for thinking that Greene et al.'s neuroscientific research gives us good reason to privilege our characteristically consequentialist intuitions over our characteristically deontological ones. (pp. 315-316, emphasis *mine*)

Even on the uncharitable interpretation, these arguments are different than the main argument. On the charitable interpretation, Greene does not make any of the above arguments, and it is merely a temptation (of argumentative opponents) to read these arguments into Greene. To be clear, I do not think that Greene makes these arguments, and I think that Berker is tempted here. When he says that they invoke these arguments for rhetorical force, he also fails to provide any relevant citations to Greene.[73] To be fair, I think that there are some arguments in the neighborhood, but these arguments are not deductions as Berker presents them. Rather, they are abductive arguments within a particular normative context that are meant to lend credence to normative conclusions. For example, consider Evolutionary History. There is an abductive argument in the neighborhood for not using judgments evolutionarily attuned to a different social environment to solve the problem of, say, anthropogenic climate change. I do not have the space to fully fill out this thought, but I trust the reader can fill it in for themselves.

So if we ought not interpret Greene as making the invalid arguments, all that is left of them is complaints that particular premises are undersupported or tendentious or that other moral philosophers disagree with some premise or needed bridging principle. But these are not logical errors or matters of shoddy inference. These are matters of substantive debate. So, in the end, there isn't really a first horn in this dilemma, at least so stated.

[73] Although he does cite Singer at one point. I agree that Singer is doing something other than what Greene is doing, argumentatively. But this is outside the scope of this paper.

4.3 Evaluating the Second Horn

Again, the second horn says that attempts to derive normative implications from the neuroscientific data rely on substantive normative assumptions that render the neuroscience normatively insignificant. Here is the argument that Berker (2009) attributes to Greene:

> **THE ARGUMENT FROM MORALLY IRRELEVANT FACTORS**
>
> P1. The emotional processing that gives rise to deontological intuitions responds to factors that make a dilemma personal rather than impersonal.
>
> P2. The factors that make a dilemma personal rather than impersonal are morally irrelevant.
>
> C1. So, the emotional processing that gives rise to deontological intuitions responds to factors that are morally irrelevant.
>
> C2. So, deontological intuitions, unlike consequentialist intuitions, do not have any genuine normative force. (p. 321)

It is worth noting at the start that Greene (2010) more-or-less accepts this characterization of his view. So the only interesting question is whether Berker has reasons sufficient to reject this argument.

Berker offers three different "worries" about this argument. The first worry is that P1 might be false. The initial characterization of the "personal" / "impersonal" distinction does not neatly track the "gives-rise-to-characteristically-deontological" / "gives-rise-to-characteristically-consequentialist" judgment distinction. The ME HURT YOU criterion seems to fall prey to Kamm's Lazy Susan Variant:

> **LAZY SUSAN VARIANT:**

> A runaway trolley is heading toward five innocent people who are seated on a giant lazy Susan. The only way to save the five people is to push the lazy Susan so that it swings the five out of the way; however, doing so will cause the lazy Susan to ram into an innocent bystander. Is it appropriate for you to push the lazy Susan so that the five people swing out of the way?

As Berker (2009) says:

> Kamm's intuition about this case is characteristically consequentialist: she thinks it is permissible to push the lazy Susan, thereby killing the one to save the five. However, in doing so one would initiate a new threat (me) that causes serious bodily harm (hurt) to a person (you), so this case counts as a personal dilemma according to Greene et al.'s criteria. (p. 311)

This is meant to show that the Argument from Morally Irrelevant Factors cannot go through as stated. We have a dissociation between the emotional processing and the property that it is purported to respond to.

This particular worry, though, doesn't seem to have much force. Or, at least, it is not going to have much force against any empiricist moral psychologists who is already metaphilosophically committed to methodological naturalism and who is an experimentalist. One way to respond is to say that this is an invitation for further refinement of the target property. This is how science operates, and there is nothing especially troubling here.

The second "worry" is closely related to the first, so I bring it up now so that I can address Worry 1 and Worry 2 with the same evidence. I quote the Worry 2 in full:

> Even if we were able to find a way of characterizing the factors which deontological judgments are responding to that makes P1 true, it is far from clear that P2 would still seem plausible. It is one thing to claim that a faculty which responds to how "up close and personal" a violation is responding to morally irrelevant features, but quite another thing to claim that a faculty which responds to whatever the sorts of features are that distinguish the footbridge case from the trolley driver case is responding to morally irrelevant features. Once we fix on what those features are, P2 may well strike us as false. (Berker 2009, p. 324)

This worry responds to my reply to the first worry. This says that even if we suitably refine P1 so that it comes out true, P2 may be false. There exist whichever sort of features distinguish Switch from Footbridge. But some of those features may not seem morally irrelevant on reflection.

The reply here is two-fold: first, this reply essentially admits that "up close and personal" as a criterion is easier to dismiss as morally irrelevant than other features. That is, it is tacitly assumed by this reply that "personalness" is either not really morally relevant in the moral-harm domain or that it is easier (relative to unnamed alternatives) to dismiss personalness as morally irrelevant. This is a major (but in my view sensible) concession to Greene, although there are some moral philosophers who would want to insist that it's morally relevant.[74]

[74] I thank Justin Bernstein and Samuel Freeman for forcing me to clarify this point.

The second reply is that this worry is a lot like the first: there could be confounds. So the reply is a lot like the first: if you think that there are confounds or that there could be confounds, then run some more experiments isolating whatever you think has been blurred or shielded! Greene actually ran some more experiments and found that people were much more likely to say "Yes" to variants of Footbridge that involved a pushing a switch to release a trap door (either up close or remotely) than to variants that involved directly pushing the person (either with hands or with a pole).[75] This puts additional, although not decisive, weight on the idea that a large number of people are preferentially responding to some property closely in the neighborhood of personalness.[76]

The final worry is that the argument is invalid. C2 does not follow from C1, unless we add a bridging premise that says that characteristically consequentialist judgments do not respond to morally irrelevant factors. More specifically, characteristically consequentialist judgments would have to be shown not to *overlook* morally relevant factors. But, as Berker correctly notes, that is exactly what non-consequentialist moral philosophers claim, and you can fill in the factor with your favorite example (separateness of persons, integrity, and so on).

Assume that this is true. Then, says Berker, the neuroscience is completely normatively insignificant. We are now arguing about which factors are morally relevant. In fact, Berker's conclusion ought to be stronger than this, for

[75] See (Greene 2010).
[76] Where that involves something resembling ME HURT YOU combine with, loosely, "touching."

psychology itself should fall out of the picture. If the point is just that we are now in a normative argument about the relevance of factors in the world, then it doesn't really matter, from such a pure moral perspective, how we are able to interact with those factors. All that matters is that those factors exists, some are relevant, and some aren't. Neuroscience falls out of the picture, but so does psychology and any empirical domain. Berker doesn't explicitly make this inference, but it is directly implied by his argument.

The reply is that if neuroscience is normatively insignificant, then so is psychology. But psychology is not. Therefore, neither is neuroscience. Unpacking this argument, let us note that it is accepted practice in most psychology departments and among most practicing psychologists that neuroscience can serve as some evidence for a psychological theory (although there is no particular need to ground every psychological theory in pure neuroscientific evidence). Greene is trying to use neuroscientific evidence to support the dual-process theory. The dual-process theory predicts that there are some cases where the emotional, affective, quick system (call it *System 1*) will overgenerate and produce incorrect answers. Unless there is a special argument that the moral domain is unlike other domains – importantly, including other normative domains – then there is no special reason to think that neuroscientific evidence has nothing to say about the relevant moral psychology.

Berker has already admitted that the neuroscientific data for the dual-process theory is impressive. If he admits that the dual-process theory is a

psychological theory and that neuroscience supports it, then he must be committed to the claim that psychology is a red herring.

However, psychology is not a red herring. Greene (2014) offers a way to think about this. If you can combine a minimal normative claim with a descriptive claim to generate a more powerful normative conclusion, then your mixed argument is not question-begging or problematic.[77] Greene offers an example: suppose you want to know whether juries in capital cases make unfair decisions. Start with the minimal moral assumption that race is an irrelevant factor for decision. Add the psychological premise that jurors decisions are affected by the race of the defendant. You generate the stronger normative conclusion that juries in capital cases make unfair decisions.

Perhaps Berker would want to respond here that "all the work" is being done by the normative premise.[78] But this cannot be correct. You cannot generate the conclusion without the descriptive premise any more than you can generate it without the normative premise. It is special pleading to deny this point. Of course, one could try to refine this point by saying that "all the *normative* work" is done by the normative premise.[79] But I am not sure what is added. If the point is that you cannot get an "ought" merely from an "is," then the reply is that Greene was never trying to do that. If the point is you need normativity to get normativity, then everyone agrees to that. If the point is that neuroscience is

[77] See also (Kumar & Campbell 2010).
[78] I thank Pierce Randal and Samuel Freeman for this point.
[79] I thank Justin Bernstein for this point.

wholly normatively insignificant, then one should specify the sense of "normative significance" such that the point is interesting.

To summarize: Berker's arguments against Greene either mischaracterize Greene's point or rely themselves on strong normative assumptions that we have reason to reject. Both horns of the dilemma amount to the same claim: either the argument is invalid or neuroscience would be wholly irrelevant. The obviously invalid arguments should not be attributed to Greene. The case that neuroscience is normatively insignificant is either false, special pleading, or lacks concrete sense.

5 Greene's Move to Moral Theory

That said, I do believe that Berker's Dilemma applies to Greene! But the question now is to locate the normative insignificance of neuroscience. Remember my claim that if neuroscience is irrelevant, then so is psychology. What we need to find is a juncture in Greene's argument where the psychology (and hence the neuroscience) is normatively insignificant. I do not believe that you find this in the earlier work. However, I do believe that you find it in *Moral Tribes*. There, Green is trying to use his prior work (as well as the work of others) to support a global moral theory. The main argument of the book is that we should all accept classical utilitarianism as the decision procedure for resolving interpersonal and intrapersonal moral conflict because moral psychology uniquely selects classical utilitarianism as the only method for resolving such

disagreement. Because the purpose of universal morality is to resolve such disagreement, classical utilitarianism is the one true universal morality.

Greene (2013) says that there are two fundamental moral problems[80]: first, the Tragedy of the Commons, and, second, the Tragedy of Commonsense Morality (pp. 14-15). "Morality" is the solution to the first tragedy, and "metamorality" is the solution to the second. I will discuss each in turn.

The Tragedy of the Commons is familiar from Garret Hardin's (1968) work and from social science. We describe as situation as being a Tragedy of the Commons when there is a social problem of cooperation created by a misalignment of narrowly conceived self-interest and collective interest. In short, everyone benefits if everyone cooperates, but everyone has an individual incentive to defect from cooperation to maximize their expected utility. According to Greene, this is the moral problem of *selfishness*, or Me and Us (2013, p. 21).

Fortunately, says Greene, Mother Nature has lent us a helping hand in the form of morality. He says: "Morality is a set of psychological adaptations that allow otherwise selfish individuals to reap the benefits of cooperation (2013, p. 23). Given this definition, it is not surprising to see Greene (2013)vwrite sentences such as "Morality evolved to enable cooperation" (p. 23). To many

[80] Interestingly, it seems like the right way to interpret Greene is that these are *moral* tragedies because each particular problem has an impact on human well-being. This will be important for my later argument that Greene is building his moral philosophy into his moral psychology in order to vindicate his moral philosophy, in a viciously and non-virtuously circular manner.

working in the humanistic tradition, this is a misunderstanding of what morality is. But, setting aside that point for now, we can see that Green offers an evolutionary functionalist account of morality. There is a specific problem of selfishness in a group, and our brains evolved over time to have a capacity for a limited form of altruism – in particular, reciprocal altruism with in-group members (Greene 2013, pp. 23-25).

However, according to Greene, there is the second tragedy, the Tragedy of Commonsense Morality. Greene hypothesizes that different groups have different moralities. These moralities regulate in-group individual selfishness by disposing people to act in a manner consistent with reciprocal altruism. But groups come into contact with one another, they have different moralities,[81] they come into conflict, and that conflict affects human well-being. Morality, then, solves the problem of the Tragedy of the Commons but creates the problem of the Tragedy of Commonsense Morality. So what we need instead is a metamorality, "a moral system that can resolve disagreements among groups with different ideals, just as ordinary, first-order morality resolves disagreements among individuals with different selfish interests" (Greene, p. 26).

I will discuss later in the diagnosis section of my paper how exactly the moral psychology – in particular the dual-process theory – is supposed to factor into this explanation. But for now notice that morality is an adaptation and

[81] It's always been unclear to me whether these are different *moralities* or different *prescriptions* from the same morality.

Greene suggests that it is perhaps fairly modular in nature.[82] Metamorality, on the other hand, is meant to rely upon some executive functioning and domain-general reasoning ability (aka, rely on System 2).

According to Greene, morality and metamorality both help to resolve particular problems. The Tragedy of the Commons, one might reasonably suppose, in endogenous to the human condition. The Tragedy of Commonsense Morality, on the other hand, is an acute problem for people in a modern age, Greene suggests. While he allows that many problems associated with the Tragedy of Commonsense Morality have been solved by technical and legal solutions, Greene (2013) highlights remaining problem of conflicting local moralities: global poverty, violent conflict, terrorism, and global warming / environmental degradation (p. 98). These are "moral conflicts" because groups with different local moralities have either different moral values or different orderings of the same moral values. However, note that each Tragedy concerns issues that impact human well-being and that require cooperative solution:

> Morality is nature's solution to the problem of cooperation within groups, enabling individuals with competing interests to live together and prosper. What we in the modern world need, then, is something like morality but one level up. We need a kind of thinking that enables groups with *conflicting moralities* to live together and prosper. In other words, we need a *metamorality*. We need a moral system that can resolve

[82] In the psychological language, morality is largely a product of System 1 processing.

> disagreements among groups with different moral ideals, just as ordinary, first-order morality resolves disagreements among individuals with different selfish interests. (Greene 2013, p. 26)

However, it is not the case that any normative theory could qualify as a metamorality. Rather, only normative theories that could in principle resolve the Tragedy of Commonsense Morality can qualify as candidates for metamorality. This restriction disallows certain sorts of normative relativist positions, such as Gilbert Harman's conventionalist ethic.[83] Other relativist positions – such as a constructivism that claims that all humans, *really*, have the same interests or moral views – could in principle be allowed but would be ruled out on the grounds that it's not true that all humans have the same interests or moral views.[84] But the main normative theories that will be candidates for metamorality are those that tend toward anti-relativism and universalism. In short, the normative theories should have the conceptual resources to diagnose and resolve disputes between competing local moralities.

Note that this still leaves everything quite indeterminate. There are, in principle, many normative theories that have the conceptual resources to diagnose and resolve disputes between competing local moralities. Here Greene introduces what I take to be the most important restriction in all of his work, a restriction that ends up, according to him, leaving only one candidate for an acceptable metamorality. I turn now to this restriction, in order to later discuss

[83] See (Harman 1975).
[84] At least at a level of grain that allows for the possibility of the Tragedy of Commonsense Morality.

how the combination of the restriction and the findings of moral psychology are meant to secure classical utilitarianism as the correct metamorality.

6 Possession Argument

One argument that Greene uses extensively is what I will call the *Possession Argument*. This argument is not fully articulated but is frequently assumed in order for him to get to his requisite conclusion. What follows is a rational reconstruction of his view.

According to the Possession Argument (PA), an acceptable global morality system must satisfy two desiderata:

> (D1) humans with basically normal psychologies must be able to comprehend what the system asks of them for each action-choice,

and

> (D2) humans with basically normal psychologies must care about or value that which the morality system requires that they care about or value.

I bundle the cognitive and motivational components together as *possession*:

> (PN1) humans must possess the psychological resources necessary to comprehend and value what the morality system requires,

and

(PN2) any morality system that individuals can't comprehend or care about properly or both is not an acceptable candidate for a global morality system.

According to Greene, deep pragmatism (or classical utilitarianism) is the only morality system that satisfies possession, and, thus, deep pragmatism is the correct global morality system. PA is an argument from elimination.

But what is the argument from elimination that deep pragmatism is the correct metamorality? There are two ways to tighten the solution space: first, we examine existing proposals for a decision procedure that would resolve the outstanding conflict and stipulate that those are the only proposals. In contemporary Anglophone normative ethics, the three primary candidates are virtue ethics, Kantian deontology, and utilitarianism. Second, we demand that the candidates encapsulate values that are shared by all because it's within the context of a shared value that conflicts can be resolved by appeals to empirical evidence. It's partly because they don't encapsulate share values that Greene concludes that virtue ethics and Kantian deontology are unsuitable candidates for metamorality.

Greene's argument against virtue ethics is quite similar to his argument against deontology. Greene argues against virtue ethics as follows:

(P1) Virtue ethics needs to appeal to the virtues to resolve conflicts.

(P2) Different tribes have different virtues.

(P3) The different virtues of the different tribes conflict.

(P4) There is no rule of priority among the virtues that settles conflict about the virtues.

(P5) Such a rule is necessary for virtue ethics to act as a decision procedure.

(C) ∴, Virtue ethics is not an appropriate metamorality.

The argument against virtue ethics is similar to the argument against deontology. Greene's argument is deliberately targeted at modern forms of deontology that stress a rights-based approach:

(P1) Rights-talk consists of evidence-insensitive moral demands.

(P2) Our automatic settings issue evidence-insensitive moral demands.

(C1) ∴, Our automatic setting issue rights-talk.

(P3) Relying on automatic settings creates or reinforces the tragedy of common-sense morality.

(C2) ∴, Rights-talk creates or reinforces the conflict.

(P4) That which creates or reinforces the conflict can't resolve the conflict.

(C3) ∴, Rights-talk can't resolve the conflict.

What's similar in these arguments is that they are both forms of the argument from disagreement.[85] The argument against virtue ethics is explicit in this regard: different tribes extol different virtues, and these different virtues conflict. Because these conceptions of virtue are not universally shared across cultures, we should be skeptical of the idea that appeals to virtue can solve the tragedy of common-sense morality. In fact, the tragedy of common-sense morality is largely a bi-product of cultures with different valuations of different character traits violently disagreeing with one another. An example of this would be debates about gender roles in society: different local moralities ascribe different character properties to their normative ideal of, say, "woman." The disagreements about the schema for 'woman' translate into intertribal conflict about how best to organize society with respect to women as well as which behaviors and attitudes are appropriate of and toward women. Disagreement about the schema for 'woman' can't be resolved by invoking a particular schema for 'woman.'

The argument against rights-oriented deontology is an implicit argument from disagreement. What's important in the argument is that while it's uniform that we all have automatic settings or processes, the outputs of these setting or outputs are not uniform and are strongly correlated with the particular local morality that we were raised in. According to the argument, automatic settings yield moral absolutes that are insensitive to evidence, and rights are a form of moral absolutes that are insensitive to evidence. Different local moralities disagree about what counts as a right. Some cultures believe that children have a

[85] See (Mackie 1990).

right not to endure any form of corporal punishment. Other cultures believe that parents not only have a right but also have an obligation to corporally punish their children. Either side invoking its supposed "right" cannot resolve this disagreement. The argument says that, by definition, when people argue using the language of moral rights and rival disputants assert "P" and "Not P," there is no evidence that could get either side to rationally concede. Automatic settings are *automatic* and, in this context, assumed to be *inflexible*.

In the next section, I will both discuss the positive argument that utilitarianism satisfies PA and present an objection side-by-side. The objection forces Greene into a dilemma that I will detail at the end of the section.

7 The Modal Objection to the Possession Argument

Now, let's point out the obvious problem with Greene's conclusion that utilitarianism is the correct metamorality: it's not clear that there is a sense of "can" such that utilitarianism satisfies PA while deontology and virtue ethics do not. Remember what PN2 tells us:

> (PN2) any morality system that individuals can't comprehend or care about properly or both is not an acceptable candidate for a global morality system.

Given what's been said, we should expect that utilitarianism is the correct metamorality. But Green does not show that. Rather, he shows that utilitarianism *may* be the correct metamorality. Here is what I understand he proves – call it *Utilitarian Metamorality*:

P1. Utilitarianism is based on three ideas: experience is what matters, and everyone's experience counts the same, and we should maximize good experiences and minimize bad experiences.

P2. Everyone cares about experience.

P3. Everyone can care about impartiality.

P4. Everyone cares about maximizing good experiences.

C1. ∴, Everyone can care about utilitarianism.

P5. It's possible to resolve the conflict by appealing to values shared by all.

C2. ∴, Utilitarianism may resolve the conflict.

Some commentary on this argument is clearly necessary. First, remember that this is an argument for utilitarianism as a decision procedure for resolving conflict between competing groups with different local moralities. Second, P2 means that people care about experience *because* it's experience, and not simply because it's *their* experience. Some may find this premise objectionable, but I allow it for the sake of argument. Third, P5 is the optimistic interpretation about the likely effects of metamorality on metamoral problems. Fourth, if it weren't already evident, I'm interpreting "cares" as "values," in the sense that if we must appeals to values shared by all, we must appeal to what all care about.

As you can likely tell, I think the action lies at P3 and how we should interpret it. But before that, we need to make sure Greene accepts P3. I think Greene accepts P3 because anything *stronger* than P3 would be both false and bad for his argument. For example,

P3*. Everyone cares about impartiality.

is clearly false, and besides being false, it's being false is crucial for the tragedy of common-sense morality to get off the ground.[86] Remember that both the tragedy of the commons and the tragedy of common-sense morality both represent a failure rooted in partiality. In the tragedy of the commons, the partiality is self-directed, and in the tragedy of common-sense morality, the partiality is directed toward the group of which I am a member. However, if

P3^. No one cares about impartiality.

then utilitarianism violates P4.[87] Moreover, if P3^, then utilitarianism fails to satisfy possession, and failing to satisfy possession is sufficient to eliminate a morality system from contention for metamorality.

Because P3^ would violate the possession requirement and because P3* is obviously false, the best that Greene can develop by way of argument is a modal

[86] The same point applies, *mutatis mutandis*, to the maximizing element of utilitarianism. I have backgrounded the maximizing element for the sake of argument, but I could run the same argument, with suitable modification and reference to behavioral economics, on that element.

[87] Another note on P3^ is that it's false because of a combination of Bishop Butler reasons and empirical facts. See (Butler 2006).

claim that people *can care* about morality. That is, the values encapsulated in utilitarianism, specifically the value of impartiality of ethical consideration, are open to humans with otherwise normal psychologies.

How are we to interpret "can?" I *can* come to like indiscriminate violence, in a certain sense of "can." It's a conceptual possibility. But in a practical or real-world context, it may also be appropriate to say that I *can't*. My character is such that there is no practical path by which I can come to like indiscriminate violence, given the set of attitudes and beliefs that I have. If "can" is interpreted in the first way, then I don't see how that helps utilitarianism or counts in its favor. We are after neither merely conceptual nor bare metaphysical possibility. For if we were, then it's surely true that it's a conceptual possibility that people *can* come to share the same virtue characterization or adherence to the same well-defined list of rights. But that implies that utilitarianism is no better off in this regard than virtue ethics and deontology, and so the argument fails to deliver the requisite conclusion.

But, in a "relative to an agent's current psychological set" sense of "can," then not everyone can care about impartiality.[88] That is, given most people's current attitudes and beliefs and given how ingrained and recalcitrant those attitudes and beliefs are, it is not true that there is a practical path by which people can come to care about impartiality. This is the well-documented case of loving relationships. There is a domain of justified partiality, most people feel,

[88] Compare to Bernard William's discussion of internal reasons in (Williams 1981a).

and considerations of impartiality don't full map out the justifications for acting on certain kinds of impartiality. *Impartiality* doesn't make me save my wife when given the choice of saving my wife or saving a stranger, nor would invocations of impartiality serve to justify the choice to my wife or others. As Williams (1981b) would say, that's "one thought too many," and it represents a distorted philosophical idealization of the substance of our ethical lives to pretend otherwise.

So Greene owes us an account of "can" such that people can get utilitarianism but not deontology or virtue ethics. I hazard to guess that such a "can" does not exist.

8 21st Century Psychology Cannot Save the Possession Argument

Greene attempts to pull the relevant "can" out of his moral psychology, in particular from his neuroscientific data. But, if you look at what he actually says, there is little support for the idea that people can care about that which utilitarianism requires that they care, in a sense of "can" that also eliminates deontology and virtue ethics from consideration. Instead, what we receive is a mish-mash of personal intuition, motivated appeals to spurious evidence, and hand waving. But, if you do not know enough about the relevant neuroscience *and* about philosophical moral theory, then it is easy to miss *both* where the argument goes wrong and *why* someone as smart as Greene would put forward arguments that are clearly bad.

Let's start with evidence that Greene attempts to answer to Modal Objection by appealing to the brain. Greene attempts to sidestep the existence of moral truth by focusing instead on the epistemology rather than the metaphysics. He writes:

> Once upon a time, I thought that this (TN: does moral truth exist?) was *the* question, but I've since changed my mind. What really matters is whether we have direct, reliable, non-question-begging access to the moral truth – a clear path through the morass – not whether moral truth exists. For the reasons given above, I'm confident we don't have this kind of access. (If there are authoritative ways to resolve moral disagreements that don't rely on divine revelation, pure reasoning, or empirical investigation, I've not heard of them.) Once we've resigned ourselves to working with the morass, the question of moral truth loses its practical importance. . . . Resigned to the morass, we've no choice but to capitalize on the values we share and seek our common currency there. (Greene 2013, pp. 188-189)

Here it is clear that Greene is searching for shared values because he believes that we do not have reliable epistemic access to the moral truth, on the hypothesis that the moral truth exists. That is, regardless of whether the moral truth exists, we still face the problems caused by the Tragedy of Commonsense Morality, and, hence, we still have a need for a metamoral solution. The metaphor of "common currency" relied upon the idea that we can translate our moral concerns (anchored in different values that are not shared) into some shared value or set of shared values that allow for explicit comparison of different choice options. This

means that, under a shared value, deliberations about what we ought to do will largely come down to figuring out what the empirical and non-moral facts are like.

If we remember the arguments against rights-based deontology and virtue ethics, then we remember that the problem with each is that the relevant values are not shared, or not shared widely enough. So, for example, different tribes extol different rights or different virtues. You can't appeal to one of those rights or virtues in order to decide which right or virtue ought to be endorsed. You need something else instead, some other value. Greene says that the shared values concern experience, impartiality, and maximization. We all care (or can care) about experience as experience. We all care (or can care) about impartiality. We all care (or can care) about maximization. Combine the three, and you have utilitarianism.

The evidence that these values are shared is supposed to come from science itself. So, while science cannot tell us what the moral truths are, says Greene, science can say which values are shared (and why) and which not (and why). So consider:

> I do not claim, however, that utilitarianism is the moral truth. Nor do I claim, more specifically, and as some readers might expect me to, that science proves that utilitarianism is the moral truth. Instead, I claim that utilitarianism becomes uniquely attractive once our moral thinking has been *objectively improved* by a scientific understanding of morality.

> (Whether this makes it the "moral truth" I leave as an open question.) Although we may not be able to establish utilitarianism as the moral truth, I believe that we can nevertheless use twenty-first-century science to vindicate nineteenth-century moral philosophy against its twentieth-century critics. (Greene 2013, p. 189)

I claim that Greene thinks that utilitarianism will be uniquely attractive because, if the explanation works, only utilitarianism satisfies PA. And he explicitly claims that *science* can show whether a particular candidate for metamorality satisfies PA. If science can show that only *one candidate* can satisfy PA, then science can "vindicate" utilitarianism in the sense that, if we agree that the Tragedy of Commonsense Morality is something to be avoided, then, given unique satisfaction of PA, we should avert to utilitarianism to solve problems related to the environment, global poverty, terrorism and the like.

So what is the scientific evidence that vindicates utilitarianism? Surprisingly, there does not exist scientific evidence that people care about experience as experience, care about impartiality, and care about maximizing good experience as such. Or, if such evidence exists, Greene does not provide it.

For the sake of argument, I accept Greene's argument that System 2 is a *maximizing* system. That is, for any value backgrounded by the system, the system will try to maximize that value. The work of Kahneman and Tversky, especially Prospect Theory, attempts to explain satisficing and loss aversion as primarily a function of System 1, such that a maximizing System 2 could, in

principle (or at least for some people), override the aversion.[89] Maximizing *per se* is not the target.

Moreover, I will not, in this paper, argue extensively against the idea that people care about experience as experience. Greene does not offer psychological or neuroscientific evidence that people value experience as experience. By "experience as experience," I mean *experience as such*. This is the familiar idea from the classical utilitarian tradition that no one's experience counts for any more than anyone else's experience and that my current experience does not count for more than my later experience simply in virtue of happening now.[90] Greene deploys traditional philosophical argumentation and intuition-pumping in order to secure this consideration. So, for example, he rehearses a familiar Aristotelian regress argument meant to show that asking "why care about that?" in relation to happiness has a quizzical or nonsensical air. But intuitions vary on this point, and plenty of people do not think the question so quizzical. Ultimately, Greene balks at extending the regress argument to secure the conclusion that all that really has value is happiness – he remains content with the idea that many chains *do* in end happiness. I am willing to spot him this conclusion.

Greene also does not offer psychological or neuroscientific evidence that people value impartiality, and I am not willing to spot him the conclusion that people care about impartiality. Nor am I willing to spot him the conclusion that

[89] See (Kahneman 2011).
[90] See (Sidgwick 1981).

people *can care* about impartiality, with some interpretation of "can" such that people *cannot* come to value what's required by deontology or virtue ethics. Again, there are two components here: people must have the conceptual resources to understand what impartiality is, and people must have the motivational resources to care about impartiality, in the sense of being able to act from direct impartial concern.

Greene goes through standard evidence that shows that most human beings are partialist and that their sympathies and altruistic concerns for others are limited in various ways. This includes evidence concerning kin altruism, direct reciprocal altruism, and indirect reciprocal altruism.[91] Limited altruism is a primary driver of the Tragedy of Commonsense Morality. So what Greene needs is some way to escape parochial altruism – some metamorality, the components of which people are able to possess.

Grant that System 2 is a maximizing system that seeks to produce optimal consequences. A good question to ask, as Greene (2013) himself notes, is "Optimal for whom" (p. 199)?

Note that you simply cannot appeal to System 2 itself to answer this question, as one who does not know much about cognitive psychology, behavioral economics, or neuroscience may want to do. Maximizing *per se* does not answer the question, and "for whom?" is an input into the system, rather than a weight of the system that operates on inputs to produce outputs. System 2

[91] For a more detailed discussion, see (Wilson, E. O., & Hölldobler, B. 2005).

is a general-purpose action planner that "is, by necessity, a very complex device that thinks not only in terms of consequences but also in terms of the trade-offs involved in choosing one action over another, based on their expected consequences, including side effects" (Greene 2013, p. 199)[92] However, we can (and do, as a matter of fact) treat deontological and virtue ethical considerations as goals (as potential end-state consequences), and we can have trade-offs between competing considerations. For example, we may want to produce a state of affairs where we act as the virtuous person would act. Accept for the sake of argument that there are multiple virtues and that the virtues are distinct and non-identical. Accept for the sake of argument that the virtuous person is the person with virtues V1, V2, . . ., Vn.[93] We may have alternative actions {A1, A2, . . ., An} to choose from. Some virtues will call for certain actions, other virtues for other actions. But there exists some determinate action that is the action that the virtuous person would do. Among that set of alternative actions is the action that the virtuous person would do, and presumably that action would involve choosing one action over others, based (at least in part) on a consideration of expected consequences, including side effects. The virtuous person could decide to act partially or to discount the considerations of at least some others in the causal wake of the action. All this is consistent with the characterization of System 2. So we need additional material to get to impartiality, for it is consistent with System 2 that it produces an ethical output that does not yield the relevant partiality.

[92] *Moral Tribes*, 199.
[93] See (Hursthouse 1996).

Greene offers two ways to get to impartiality, but neither has much to do with relevant psychological or neuroscientific research. The first concerns selfish individual actors deciding how to split a pot. Say that there are ten actors and one-thousand coins. No actor has a threat advantage over any other actor. How to split? The solution, says Greene, is an equal split, for there are no power asymmetries that could give rise to a motive to defect from equal split. The solution is "stable," in the sense that no individual actor has an incentive to defect from equal splitting.

But notice that this is just game-theoretic analysis and has nothing in particular to do with how actual human beings make their decisions. Of course, if people were trying to maximize their own well-being and if there were no power asymmetries, then people would have the conceptual and motivational resources necessary for them to "get" impartiality. But we already know that people and the world are not like that: people are not merely selfish (the evidence of limited altruism proves that) and power asymmetries have always existed (anthropology, history and political awareness proves that). How do creatures *like us* who live in conditions *such as ours* come to "get" impartiality?

Greene tries another to get to impartiality in another way by drawing on one of the central ideas of Peter Singer's *The Expanding Circle* (1981). Greene starts this time with human agents who are predominantly egoist.[94] People care for themselves, for their families, for the friends, and for relevant in-group members.

[94] A much more faithful model for actual human agents.

I quote to show that there is no relevant psychological or neuroscientific data that backs up the move:

> People, for the most part, don't care very much about complete strangers. But at the same time, people *may come* to appreciate the following fact: Other people are, more or less, just like them. They, too, care most of all about themselves, their family members, their friends, and so on. Eventually, people *may make* a cognitive leap, or a set of cognitive leaps, culminating in a thought like this: "To me, I'm special. But other people see themselves as special just as I do. Therefore, I'm not really special, because even if I'm special, I'm not especially special. There is nothing that makes my interests *objectively more important* that the interests of others." (Greene 2013, p. 200, emphasis *mine*.)

This is a hand-waving explanation of a crucial component of utilitarianism. Remember that science was supposed to help pick out utilitarianism as a particular attractive metamorality. But, when we get to the important part of people "getting" all the components of utilitarianism, science exits, and magic comes in.

The first leap comes when people appreciate that others care about the things that they themselves care about. Even here, we have it that people *may come* to appreciate this fact. Of course, they may not. This is merely to throw us back again on the problem of the Modal Objection. What is the sense of "may come" such that people *will not and cannot come* to appreciate that which

deontology or virtue ethics requires that they appreciate? Second, there is no compelling scientific evidence that I am aware that suggests that most people do in fact appreciate that fact or that they may come to. What we have here is either some sort of optimism in people (which would be surprising given some of the things that Greene has said about the folk in other parts of *Moral Tribes*), or some sort of importing of moral content into the psychological explanation. That is, Greene needs it to be the case that people can cognitively and motivationally get the idea of impartiality. But he should be committed to the idea that we can say that people can cognitively and motivationally get the idea of impartiality on the basis of reliable scientific evidence. This is psychological speculation of exactly the sort that Greene deplores humanists as engaging in.

The second leap (or series of leaps) comes when people move from the idea that everyone basically cares about the same things to the incredible ideas found in the imagined monologue at the end of the quote. People move from the first idea to "To me, I'm special" to "To each person, he/she/they are special" to concluding that "I'm special, but not especially special" to "there is nothing that makes my interests objectively more important that the interests of others." There are a couple of striking things here.

First, again, people "may make" the additional series of leaps *eventually*. There is the already stated problem concerning "may make." But "eventually" is another problem altogether. How much time are we allowed to grant for "eventually" to have purchase for utilitarianism but not for deontology and virtue ethics? Here, the relevant sense of "can" is not one relevant to choice at the

current moment. If we allow that people eventually may make such additional leaps, then why can we not allow that people may make such additional leaps as is required of the other theories?

Second, the whole line of thought is just one massive non-sequitur and is invalid at each step. If people reason in such a way, then this cannot possibly be rational evidence of utilitarianism or the component under consideration. I am inclined to think that if *this* is how people get to "get" impartiality, then we have a debunking argument against the idea of impartiality, based upon its improper etiology. Invalidly drawn conclusions may be true or false, but you need some other chain of reasoning or evidence to establish them as true or false. The story given should not affect your antecedent commitments.

Third, you cannot even get to impartiality from these considerations alone. The selfish nihilist can accept the conclusion of the reasoning but not accept impartiality. The nihilist is one who thinks that nothing is objectively more important than anything else. But that does not imply that the nihilist does not find some things more important. If selfish, then the nihilist privileges his interests over others while also holding that there is no objective basis for him to do so. The same is true, with appropriate modifications, for certain relativist and subjectivist views.

In fairness to Greene, he acknowledges this. However, the answer that he gives to this objection is unsatisfying:

> But it seems that, somehow, we do manage to translate this intellectual insight into a preference, however weak, for genuine impartiality. I suspect that this translation has something to do with *empathy*, the ability to feel what others feel. Human empathy is fickle and limited, but our capacity for empathy may provide an emotional seed that, when watered by reasoning, flowers into the ideal of impartial morality. (2013, p. 201)

This will not work, though. First, a "weak preference" for genuine impartiality is not enough. What is required is that people have an overriding preference for genuine impartiality in areas of metamoral concern. If people merely have a weak preference, then that preference will be trumped by other preferences in the relevant profiles. But that is just to say that people will not, in fact, be motivated to act on the idea of impartiality, which is just to say that utilitarianism fails to satisfy PA. Second, what does it mean to be "watered by reasoning?" The metaphor is inapt, for Greene holds that the relevant kind of reasoning is the sort of stuff that System 2 does. But we have already said that System 2 is a maximizing-relative-to-a-value system. There is no obvious or perhaps foreseeable path from "Maximize the pleasurable experiences of those to whom I have concern" to "Maximize the pleasurable experiences of all people." That is exactly what is at stake.

Greene (2013) ultimately ends up admitting that he has no idea how the idea of impartiality came about in humans with the sorts of brains they have. But, he says,

> I'm fairly confident of two things. First, the ideal of impartiality has taken hold in us (we who are in on this conversation) not as an overriding ideal but as one that we can appreciate. None of us lives perfectly by the Golden Rule, but we all at least "get" it. Second, I'm confident that the moral ideal of impartiality is a manual-mode phenomenon. This ideal almost certainly has origins in automatic settings, in feelings of concern for others, but our moral emotions are themselves nowhere near impartial. Only a creature with a manual mode can grasp the ideal of impartiality. (p. 201)

Note that this is only a possible explanation of the cognitive grasp of impartiality, not the motivational grasp. An overriding ideal, on this construal, would be one that would motivate us at each choice-point. But we are not so motivated. Rather, we "get" impartiality. I see no way to read this passage that says anything above that we have the conceptual tools to understand what impartiality requires of us. Notice again that there is no psychological or neuroscientific evidence that is brought to bear on this question. Rather, there is Greene's confidence that anyone who has to capacity to ask the metamoral question has the conceptual resources to understand impartiality. This confidence is less than the advertised standard of proof Greene offered before. Moreover, there is again the idea that impartiality (or genuine impartiality, extending to all) comes about from System 2 operations. But given everything said about System 2 above, we have absolutely no reason to accept that

characterization. Moreover, given everything that Greene has said about System 2, it is completely unclear what the scientific evidential basis is for his confidence.

We still have the lingering motivational component. Greene offers an unclear analogy to explain the motivational component. There is a difference between shopping for food while hungry and while full. We can be motivated to shop for food based entirely on automatic settings. But we can also be motivated to shop for food while completely full. Even though shopping while full will have something to do with your automatic settings, it does not fully rely on your automatic settings. Your manual settings can allow you to shop for things that you do not desire at all presently, on the basis that you can project into the future what you would like. You can also shop for other people's food. If you relied merely on automatic settings, then you would just get the things you like. The moral, at the end of this confusing metaphor:

> Somehow, the human brain can take values that originate with automatic settings and translate them into motivational states that are susceptible to the influence of explicit reasoning and quantitative manipulations. We don't know exactly how it works, but it clearly does. (Greene 2013, p. 202)

Appeals to brute fact are surprising here. The whole point was that science was supposed to show why utilitarianism was particularly attractive as a metamorality. But how does this metaphor even work? We have problems dealing with individual selfishness. Limited altruism helps solve some of the problems associated with individual selfishness. But we also have problems of

group selfishness. So System 2 just hijacks the output of System 1 (limited altruistic concern) and translates that output into some kind of motivational state such that people either care about or can care about genuine or full impartiality. This is not a genuine scientific explanation. This is pounding on the table in order to continue to secure the conclusion that utilitarianism is the correct metamorality.

I conclude this section by underlining the fact that Greene has not established his case as he said he did. He said that science would show that utilitarianism is the most attractive candidate for metamorality by showing how utilitarianism satisfies PA. But, when we get to impartiality, Greene offers a mish-mash of motivated appeals to evidence, intuition-pumping, motivated appeals to brute facts, and so on. If we pay attention to how humans act and if we pay attention to relevant features of System 2, we see that there is no way that Greene was able to pull the rabbit out of the hat with the resources that he has.

9 Conclusion

In this paper, I have diagnosed different problems with the moral psychology offered by Joshua Greene. I started with showing that the objection that most humanists take as decisive – Berker's Dilemma – actually lacks force because it is involved in exactly the sort of game of smuggling normative content that it accuses Greene of and because it grants to Greene too many of the descriptive components for those normative complaints to have any upshot. Then I turned my attention to the particular normative claims that Greene wants

to make. I exposed how those normative claims turn on how we are to understand the Possession Argument: that human beings with otherwise normal psychologies must have the cognitive and motivational resources necessary to "get" what the particular candidate for metamorality says they must get. One way to understand this point is to grasp that any particular normative theory will bring with it some empirical or psychological commitments as well. If we can show that those empirical or psychological commitments cannot be cashed out satisfactorily, then we have provided at least some reason to reject the normative theory under consideration.

Greene thought that he could show that people can "get" utilitarianism but not "get" rights-based deontology or virtue ethics. What I have attempted to do is force Greene into a dilemma:

> (D1) Either people have the cognitive and motivational resources to "get" utilitarianism and also deontology and virtue-ethics, or
>
> (D2) People do not have the cognitive and motivational resources to get deontology and virtue ethics and also utilitarianism.

My support for the dilemma came about by closely examining what sort of evidence that Greene provided to show that people "get" utilitarianism. I provided the Modal Argument to show that it cannot be the case for Greene that people, right now, in fact, "get" utilitarianism in the right way. If they did, then there would be no issue of metamorality. Instead, at best, people "can get" utilitarianism. We can revise the dilemma:

(D1*) Either people have the cognitive and motivational resources such that they can "get" utilitarianism and also deontology and virtue ethics, or

(D2*) People do not have the cognitive and motivational resources such that they can "get" deontology and virtue ethics and also utilitarianism.

But Greene still requires evidence that he can get out of the dilemma, that people can "get" utilitarianism but not potential metamoral rivals. I argued that he thinks that science will provide us with the relevant sense of "can." There is something about the brain, he thinks, that makes utilitarianism quite attractive.

But when we looked at the evidence, we noticed that he admits that System 2 is a general-purpose maximizing-relative-to-a-value system. So there has to be some other way for people to come to cognitively and motivationally grasp utilitarianism. First, Greene offered a game-theoretic story involving pure egoists with equal threat advantage. This is a thought experiment and does not rise to the level of psychological evidence in the relevant way (we are after descriptive models, not normative models). Second, Greene appealed to Singer's idea of the expanding circle. But Greene's version of the reasoning behind the expansion is invalid and lends no rational weight to utilitarianism. Moreover, there is no explanation of the mechanism by which one comes to full impartiality from that reasoning, for the reasoning does not end in full impartiality (just that nothing objectively matters). Finally, we saw Green appeal in both the cognitive and motivational cases to brute facts, but he is not entitled to that appeal (and by his own evidential standards).

I have made the kinds of arguments that I have made by appealing to a wide range of considerations, including game theory, social science, cognitive science and ethical theory. But the general thrust of my argument can be reduced to the following: Greene is an ardent utilitarian who would like others to also be utilitarians. From this, he engineers a moral psychology that is meant to show that utilitarianism is the most attractive candidate for universal morality. This kind of circularity is bad.

The arguments that I have provided have attempted to identify and eliminate excesses of moral content in Greene's psychological explanation. I have attempted to draw from diverse evidence bases – importantly, including social science, cognitive science, and ethical theory – to show how Greene's moral psychology is really in the service of his utilitarian ethic. The moral psychology itself, and the evidence offered for it, provides insufficient rational support for utilitarianism in the end. But, when we look at how Green ignores what he acknowledges to be the case about System 2 in order to secure claims about the possession of impartiality, we can see that moral content is infecting his psychological explanation. He knows that he needs a psychological explanation, but, in the end, we are not given one. We have appeals to intuition, to brute fact, and to features of System 2. But, given what we know of System 2 and what Greene admits, System 2 could never, by itself, transform limited altruism into full impartiality. That is what Greene needs for his arguments, so that is what he claims.

The moral of this paper, then, is that we should be wary of those with a normative agenda who are also doing psychology. In particular, we should be wary of psychological explanations that are just so convenient for securing some determinate normative conclusion. The suspicion is that, when you examine the arguments and evidence concerning psychologies uniquely picking our normative theories, the normative theory has already been illicitly imported at some step. The importing need not take the form of putting some determinate moral principle in the head; rather, you can have the import of moral content into psychological explanation when the psychologist is offering a purportedly neutral psychological characterization that *just so happens* to have some particular normative upshot. When some moral psychologist or philosophers introduces some psychological mechanism or explanation just to secure a normative conclusion, you should nearly always reject the mechanism or explanation. But this requires some understanding of the relevant normative considerations, and normative theory, at play.

Bibliography

Berker, S. (2009). The normative insignificance of neuroscience. Philosophy & Public Affairs, 37(4), 293-329.

Berker, S. (2014). "Does Evolutionary Psychology Show That Normativity is Mind-Dependent?" in D'Arms, J., & Jacobson, D. (Eds.). (2014). *Moral Psychology and Human Agency: Philosophical Essays on the Science of Ethics*. Oxford University Press.

Butler, J. (2006). *The Works of Bishop Butler* (Vol. 14). Boydell & Brewer.

Davidson, Donald (1984) [1974]. "Ch. 13: On the Very Idea of a Conceptual Scheme". *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Farah, M.J. (2010). Neuroethics: An Overview. In M. Farah (Ed.), *Neuroethics, An Introduction with Readings*. Cambridge, MA: MIT Press, 2010.

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

Greene, J. (2003). From neural 'is' to moral 'ought': what are the moral implications of neuroscientific moral psychology?. *Nature Reviews Neuroscience*, *4*(10), 846-850

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in cognitive sciences*, *6*(12), 517-523

Greene, J. D. (2007). The Secret Joke of Kant's Soul, in edited by Sinnott-Armstrong, W.(ed.), Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development.

Greene, J. (2010) "Notes on 'The Normative Insignificance of Neuroscience' by Selim Berker." Unpublished manuscript.f

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

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

Hardin, G. (1968). The tragedy of the commons. *science*, *162*(3859), 1243-1248.

Harman, G. (1975). Moral relativism defended. *The Philosophical Review*, *84*(1), 3-22.

Hursthouse, R. (1996). "Normative Virtue Ethics," from Roger Crisp, ed., *How Should One Live*? Oxford University Press. 19-33.

Kamm, F. M. (1989) "Harming Some to Save Others", 57 *Philosophical Studies* 227-60.

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgement. *Mind & language*, 25(5), 561-582.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kleingeld, P. (2014). Debunking Confabulation: Emotions and the Significance of Empirical Psychology for Kantian Ethics. In *Kant on Emotion and Value* (pp. 146-165). Palgrave Macmillan, London.

Kumar, V & Campbell , R. (2012): On the normative significance of experimental moral psychology, *Philosophical Psychology*, 25:3, 311-330

Lott, M. (2016). Moral Implications from Cognitive (Neuro) Science? No Clear Route. *Ethics*, *127*(1), 241-256.

Mackie, J. (1990). *Ethics: Inventing right and wrong*. Penguin UK.

Sidgwick, H. (1981). *The methods of ethics*. Hackett Publishing.

Singer, P. (1981). *The expanding circle*. Oxford: Clarendon Press.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204-217.

Thomson, J. J. (1984). The trolley problem. *Yale LJ*, *94*, 1395.

Wielenberg, E. J. (2014). *Robust ethics: The metaphysics and epistemology of godless normative realism*. OUP Oxford.

Williams, B. (1981a). "Internal and External Reasons." In *Moral luck: philosophical papers 1973-1980*. Cambridge University Press.

Williams, B. (1981b). "Moral Luck." In *Moral luck: philosophical papers 1973-1980*. Cambridge University Press.

Wilson, E. O., & Hölldobler, B. (2005). Eusociality: origin and consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38), 13367–13371.