Factlessness & Faultlessness: Individual Differences & Dimensions of Philosophical Dispute

Geoffrey Scott Holtzman

How does access to this work benefit you? Let us know!

Follow this and additional works at: https://academicworks.cuny.edu/gc_etds

Part of the Cognitive Psychology Commons, and the Philosophy Commons

Recommended Citation

https://academicworks.cuny.edu/gc_etds/431

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
Factlessness & Faultlessness:
Individual Differences & Dimensions of Philosophical Dispute

by

Geoffrey S. Holtzman

A dissertation submitted to the Graduate Faculty in Philosophy for the degree of
Doctor of Philosophy
The Graduate Center, City University of New York
2014
This manuscript has been read and accepted for the
Graduate Faculty in Philosophy
in satisfaction of the dissertation requirement for the degree of
Doctor of Philosophy.

John Greenwood

Date

Chair of Examining Committee

John Greenwood

Date

Executive Officer

Jesse Prinz

Hagop Sarkissian

Jennifer Mangels
Supervisory Committee

The City University of New York
Abstract

Factlessness & Faultlessness:
Individual Differences & Dimensions of Philosophical Dispute

by

Geoffrey S. Holtzman

Advisor: Jesse Prinz

This project addresses the question of why philosophical disputes persist, and tackles the problem of how we might better approach them. I demonstrate empirically several ways in which personality, gender, and other factors are associated with specific philosophical beliefs. Typically, one might assume that these individual difference factors are irrelevant to philosophy, and can only serve to bias philosophical disputants. Against this view, I present four case studies, which collectively highlight the different ways in which individual differences in lived experience may be inseparable from philosophical concepts themselves.
In completing this project, I am indebted foremost to Jesse Prinz and Jennifer Mangels for their kindness, time, attention, patience, expertise, training, and guidance. If not for Jennifer, I would be a charlatan; if not for Jesse, I would have dropped out of school long before I had the chance to become even that.

For teaching me how to work and how to think, and also spending about a gazillion dollars on my education, and for inexplicably reading and commenting on this entire document, I am extremely thankful to my parents, Audrey and William Holtzman.

The project benefitted enormously from critical feedback from John Greenwood, Joshua Knobe, and Hagop Sarkissian, to whom I express my gratitude.

Valuable feedback on manuscripts that would become a part of one or more specific chapters, and for that feedback I am deeply appreciative of the time and energy put forward by Lisa Aspinwall, Teneille Brown, Gina Campelia, Noel Carroll, Edward Cokely, Anne Cross, Michael Devitt, Adam Feltz, Michael Levin, Danielle Martineck, Barbara Montero, Eddy Nahmias, David Rosenthal, Jim Tabery, and Ron Whiteman.

Constructive feedback on manuscripts that have become part of this project was provided by journal editors John Banja and William Bechtel, and (especially) Graham Horswell, and to more blind reviewers than I care to admit.
For the attention they brought to the ideas and experiments in Chapter Two, I wish to thank Wesley Buckwalter and Steven Stich.

Chapter Five would not have been conceived had I not been able to study gender differences in data shared with me Fiery Cushman, Joshua Greene, Shaun Nichols, and Deena Weisberg.

Some of the arguments presented here were heavily influenced by conversations with and objections leveled by Amanda Bryant, Michael Fisher, and Ian Olasov.

For helping me make it to the beginning of this project (not to mention helping me make it to the end), my appreciation extends to Mark Alfano, Noel Carroll, Rosemarie Ianuzzi, Brian Leiter, Barbara Montero, Thomas Nadelhoffer, and Iakovos Vasiliou.
# Table of Contents

**CHAPTER ONE: INDIVIDUAL DIFFERENCES, SOUNDNESS, AND VALIDITY**  

1.1. Diagnosis & Prescription in Analytic Philosophy  
   1.1.2. Faultiness: The rationality or irrationality of inferences  

1.2. Criticisms of the Analytic Project  
   1.2.1. Factive vs. factless disputes  
   1.2.1. Faulty vs. faultless disputes  

1.3. The Taxonomic Project  
   1.3.1. Chapter Two: Pluralist relativism about factless, faultless disputes  
   1.3.2. Chapter Three: Monist realism about factive, faulty disputes  
   1.3.3. Chapter Four: Pluralist realism about factless, faulty disputes  
   1.3.4. Chapter Five: Monist relativism about factive, faultless disputes  

**CHAPTER TWO: PERSONALITY EFFECTS**  

2.1. The Nature of Philosophical Debate  
   2.1.1. Experimental Philosophy  
   2.1.2. The Big Five Personality Traits  

2.2. Method  

2.3. Results  
   2.3.1. Dualism: Could a Robot Feel Love?  
   2.3.2. Descriptivism: Are Names Identical With Descriptions?  
   2.3.3. Embodiment: Could a Brain Think Without Being Connected to a Body?
2.4. Discussion: Do Personality Effects Mean Philosophy is Intrinsically Subjective?

2.4.1. The Connection Between Personality Traits and Subjectivity

2.4.2. The Connection Between Philosophical Belief and Philosophy

2.4.3. Objection: The Discovered Effects Are Philosophically Uninteresting

2.4.4. Objection: The Big Five Do Not Truly Assess “Personality”

2.4.5. Objection: Personality Does Not Affect Philosophical Belief

2.4.6. Objection: The Philosophical Prompts Were Not Worded Fairly

2.4.7. Conclusion: Some Parts of Philosophy Are Subjective

2.5. Implication: Some Parts of Philosophy Are Not Subjective

Chapter Three: Gender Effects

3.1. Rejecting Beliefs, or Rejecting Believers?

3.2. Focus on the Expert/Gender Assumption

3.2.1. Rehashing the argument

3.2.2. Revisiting the expert/gender assumption

3.2.3. Reframing the discussion

3.2.4. Reanalyzing the data

3.3. Rethinking the moral of the story

3.4. Building on the results

Chapter Four: Conceptual Convergence

4.1. A Puzzle of Moral-Agentive Cognition

4.2. Motivating the Concept of Agency & the Concept of Motivated Agency
4.2.1. Topical and Methodological Purview

4.3. A Puzzle of Moral-Agentive Cognition

4.3.1. Type and Token Thought Experiments

4.3.2. The Source of the Puzzlement

4.4. Normative and Deflationist Accounts of the Puzzle

4.4.1. Shortcomings of the Normative and Deflationist Approaches

4.4.2. The Need for A New Approach

4.5. Paradigms of Moral-Agentive Relata

4.5.1. The Identity Paradigm

4.5.2. The Common Causes Paradigm

4.5.3. The Agency-First Paradigm

4.5.4. The Agency-Last Paradigm

4.5.5. From Psychological Models to Conceptual Paradigms

4.6. Experiment

4.6.1. Participants

4.6.2. Methods and Materials

4.6.3. Results

4.7. Discussion

4.7.1. From where do we derive the apperception of free choice?

4.7.2. What are the common conceptual schemata that allow us to coherently debate and discuss such opaque phenomena as liberty, willfulness, and agency?

4.7.3. What do people really believe about the compatibility of free will and causal determinism, and are their beliefs contradictory?
4.7.4. Why do we tend to view others as autonomous agents in proportion to the extent to which they have ‘done wrong’?

4.7.5. Why do philosophers care about free will?

4.8. Agency-Last Perspectives on Classic Metaethical Puzzles

4.8.1. Contempt, Coercion, and Control

4.8.2. The Puzzle of Moral-Agentive Cognition

4.8.3. Free Will as Moral Ether

CHAPTER FIVE: ADJUDICATING ADJUDICATION

5.1. The Double-Edged Sword

5.2. Scientific & Philosophical Skepticism

5.3. Experiment

5.4. Discussion

CHAPTER SIX: INDIVIDUAL DIFFERENCES & PHILOSOPHICAL BELIEFS

6.1. Recommendations

6.2. Shortcomings & Future directions

6.2.1. Expanding the agency-last paradigm

6.2.2. Better adjudicating adjudication

6.3. Conclusion

FIGURES

REFERENCES
List of Tables

Correlations Between Personality Factors and Philosophical Beliefs
Table 1…………………………………………………………………………….Section 2.3.

Could a Robot Feel Love?
Table 2…………………………………………………………………………….Section 2.3.1.

Are Names Identical With Descriptions?
Table 3…………………………………………………………………………….Section 2.3.2.

Could a Brain Think Without Being Connected to a Body?
Table 4…………………………………………………………………………….Section 2.3.3.

Hypothesis Evaluation: Fit Between Theoretic Paradigms and Observed Attributions
Table 5…………………………………………………………………………….Section 4.6.3.
List of Figures

Gender: Free will is compatible with determinism
Figure 1........................................................................................................Section 3.2.4.

Gender: Taste qualia might be scientifically reducible
Figure 2........................................................................................................Section 3.2.4.

Gender: A robot could never experience love
Figure 3........................................................................................................Section 3.2.4.

Gender: Minimax is no fairer than maximizing average utility
Figure 4........................................................................................................Section 3.2.4.

Gender: An unembodied brain might be able to think
Figure 5........................................................................................................Section 3.2.4.

Gender: Flipping a switch to save 5 by sacrificing 1 is wrong
Figure 6........................................................................................................Section 3.2.4.

Gender: Names refer by description, not causal history
Figure 7........................................................................................................Section 3.2.4.
Gender: Mistakenly justified true beliefs are still knowledge

Figure 8........................................................................................................Section 3.2.4.

Gender: The body, not the mind, determines identity

Figure 9........................................................................................................Section 3.2.4.

Model Specification: Theoretic Paradigms of Moral-Agentive Relata

Figure 10......................................................................................................Section 4.5.

Parameter Estimation: Agency-Last Mediation Analysis of Moral-Agentive Attributions

Figure 11......................................................................................................Section 4.6.3.

Influence of Additional Biomechanism on Sentencing by Male and Female Judges

Figure 12......................................................................................................Section 5.3.
CHAPTER ONE

TRUE PREMISES, VALID INFERENCES, AND INDIVIDUAL DIFFERENCES

Why do people disagree about the answers to philosophical questions? I take this to be the fundamental question of metaphilosophy. Only by identifying the contours of philosophical disputes, and the factors that draw people across those fault lines, can we best understand the concepts and theories in dispute. This may sound like a truism, but when more fully unpacked, it suggests a marked shift from the more traditional method of immediate reflection upon individual philosophical questions and theses. Notably, if there turns out to be a plurality of causes for the diversity of philosophical belief, then no single method of inquiry or dialogue will be sufficient for scrutinizing the entirety of the philosophical landscape. Instead, a variety of approaches will be called for. But in order to know what approach to take to a given philosophical questions—that is, what kind of answer is to be sought after—we need a taxonomy of causes of philosophical dispute. A taxonomy of this kind could help clarify kinds of philosophical problems, and modes of resolution. My purpose here is to develop, defend, deploy, and demonstrate such a taxonomy.

In Chapter One, I lay the groundwork for my approach in three steps. First, I discuss the standard approach to resolving theoretic disputes in philosophy. This approach represents a kind of methodological monism, as there is assumed to be only one acceptable form and content for the arguments given by disputants, and all others are thought to be wrong (§1.2.). Second, I make the case that a different approach—a methodological pluralism—is demanded by the nature of philosophical problems (§1.3.). Third (§1.4.), I develop and defend a 2x2 taxonomy of philosophical divergence, which is meant as a first pass at formalizing this methodological
pluralism. Fourth (§1.5.), I discuss how one would apply the approach I advocate. In Chapters Two through Five, I apply this taxonomy to a series of questions in philosophy, and in particular in the philosophy of moral agency, in order to illustrate its philosophical utility and to demonstrate its explanatory and predictive value. Chapter Six returns to the abstract issues discussed in this first chapter, rooting them in knowledge accumulated in the pages between the two.

1.1. Diagnosis & Prescription in Analytic Philosophy

How can we resolve philosophical disagreements in cases of seeming theoretic stalemate? The first step is to identify the factors that cause people to hold different philosophical views in the first place. No earnest attempt to solve a problem can be made without a prior effort to diagnose the exact nature of the problem; diagnosis precedes prescription. Traditionally, unsound philosophical arguments are diagnosed according to two dimensions. I describe them here, then go on to explain why they are insufficient to capture the causes of, or potential resolutions to, many philosophical disputes.

1.1.1. Facticity: The truth or falsity of premises

Hume divided the realm of human inquiry into two mutually exclusive kinds: “matters of fact,” and “relations of ideas” (1740/2003). Modern philosophy, which emerged in the 16th century, was defined by a radical shift in the way we establish the former. This shift led philosophers away from reliance on the Bible, ancient philosophy, and empirical observation, and toward the
use of intuitive premises. Today, the rejection of empirical premises may still be the defining characteristic of much philosophy.

It may be from Descartes that analytic philosophers derive their peculiar use of the term ‘intuition.’ While the definition and role of intuition in contemporary philosophy is subject to debate (Cappelen, 2013), for the purposes of this project I will define the term negatively and in accordance with Descartes. Here, I will use the term ‘intuition’ to mean ‘insensate apperception,’ in strict contrast with sensory perception:

Its perception is neither an act of vision, nor of touch, nor of imagination, and has never been such although it may have appeared formerly to be so, but only an intuition of the mind.

In both modern and analytic philosophy, it is typical to say that an intuition is right if it leads us to embrace a fact that turns out to be true, and wrong if it leads us to embrace a premise that is actually false.

1.1.2. Faultiness: The rationality or irrationality of inferences

Analytic philosophy, like modern philosophy, is distinguished from science by its use of intuitive rather than observational evidence to establish matters of fact. Analytic philosophy further distinguishes itself from modern philosophy primarily by formalizing the conception of the “relations between ideas,” by which we identify fault. In 1879, Frege published the first fully axiomatized system for analyzing complex sentences (1879/1931). By 1906, Russell had proven
that the completeness of such a system of predicate calculus could be achieved with just a single binary operator, negation, applicable equally to both predicates and the objects of their predication (1906). These advancements paved the way for a more formal method of philosophical inquiry to come into fashion (primarily in the English-speaking world) at the turn of the 20th century. For this reason, Frege and Russell are considered the forefathers of analytic philosophy.

Analytic philosophers typically seek to adjudicate theoretic deadlocks through the deductive elimination of all and only those arguments that are unsound, invalid, or both. Theses whose only support comes from such arguments are expected to fall by the wayside, until just a single sound argument yields a guaranteed conclusion that can be validly inferred from true premises. Faulty reasoning (which undermines both validity and soundness) and false premises (which undermine soundness), it is usually assumed, are the two causes of flawed arguments whose conclusions are not guaranteed, and which therefore fail to provide the justification necessary for knowledge. Because two arguments are only guaranteed to converge on the same conclusion if each is flawless, philosophical disagreements are generally thought to arise because one or more disputants has drawn her conclusions on the basis of faulty reasoning, false premises, or both. But this traditional explanation of why disputants disagree may be overly simplistic.

1.2. Criticisms of the Analytic Project

Can philosophical disputes always be adjudicated by appeal to universally accessible facts and decisive proof of fault? Analytic philosophy assumes that we can answer this
question in the affirmative. But against this immediate, analytic approach to theoretic adjudication, I offer a number of criticisms, arguing that the underlying causes for the diversity of philosophical intuition are, themselves, diverse. The forms of this diversity arise from two oft-overlooked sets of contrastive features of philosophical disputes.

1.2.1. Factive vs. factless disputes

Analytic philosophers typically pride themselves on the abolishment of false premises from their arguments, but the grounds for this pride remain in dispute. Many discontents within analytic philosophy believe that philosophers’ efforts to eradicate false premises from their arguments (by employing intuition rather than observation) have had unintended and deeply problematic consequences. One question these discontents raise is whether all arguments in traditional analytic philosophy actually operate on exclusively (Gibbard and Blackburn, 1992) or even any evidential (true, or even false) premises at all (Ayer, 1952; Hare, 1952;).

Are the intuitions that divide philosophical disputants always mutually truth-apt “matters of fact,” portraying either truths or falsehoods about some single set of facts, or are some such premises factless? Such cases often seem to arise when the evidentiary basis for an argument is not or cannot be (mutually) agreed upon by theoretic disputants. Both factive and factless disputes can be further distinguished into two kinds each.

1.2.1. Faulty vs. faultless disputes

The orthodox view has it that all arguments can be categorized—regardless of the truth of their
premises—as either rational or irrational. An argument that assumes some “relations between ideas” without logical warrant is invalid, and therefore fails to support its conclusion. All other arguments are valid, and if they are also sound—if their premises are also all true—then they guarantee the truth of their conclusions. A belief is only rational if it is based on valid arguments, where beliefs derived from invalid arguments—even though they may turn out to be right—are irrational. But again, we might question whether this is really the case, or whether there might be some philosophical arguments do not fall into either category.

Such cases might arise, for instance, when the frame of reference for an argument is not or cannot be mutually agreed upon by theoretic disputants. One example where this seems to have occurred is in the debate among mathematicians (Arbesman and Strogatz, 2008; McCotter, 2008; Rockoff and Yates 2009) as to whether Joe DiMaggio’s 56-game hitting streak was “so many standard deviations above the expected distribution that it should not have occurred at all” (Gould, 1988). Other cases might arise when a person satisfies the “state-requirements” for rationality, despite failing to meet the “process-requirements,” as in the hypothetical case of a man who resolves a conflict between inconsistent beliefs by dropping one due to an electric shock (Kolodny, 2005). In cases like these, philosophical disputes may simply be arational, capable of accommodating two or more incompatible modes of resolution, each of which is nevertheless faultless.

1.3. The Taxonomic Project

These contrastive characteristics—factiveness vs. factlessness, and faultiness vs. faultlessness—interact to suggest a taxonomy of philosophical questions. The application of this taxonomy is
Factlessness & Faultlessness demonstrated in Chapters 2-5. Each of these chapters focuses, at least in part, on a different aspect of philosophical cognition. However, all of these chapters focus, at least partly, on what has traditionally been conceived of as a single philosophical question. This is the question of whether agents ever deserve to be blamed for choosing immoral courses of action, if their participation in those courses of action was already determined by external forces prior to their making any choices at all. Thus, while these chapters can be read individually, they may also be read as a unified whole. By focusing on different taxonomic kinds of philosophical disputes we engage in about this putatively singular question, I show that it is in fact comprised of many disputes.

1.3.1. Chapter Two: Factless, faultless disputes

Concepts that are rooted in phenomenological experience seem especially likely to be fundamentally informed by sensory experience that cannot be shared with others. In Chapter Two, I demonstrate this point by identifying a number of cases in which the views of professional philosophers are predicted by their personalities. Philosophers who are given to worry and moodiness are more likely than their colleagues to deny that a robot could ever experience love; philosophers who have a penchant for the arts are more likely than their colleagues to say that a brain cannot think unless it is connected to a body. The contours for defining central pieces of evidence for these views—what it means to think, and what it means to love—are rooted in first-personal experiences whose factual denial by a third party is simply impossible. Furthermore, if two interlocutors are unwittingly discussing different concepts from each other, it is possible for them to both have internally consistent views, even if they do not
draw the same conclusions as one another.

1.3.2. Chapter Three: Factive, faulty disputes

There are, of course, many cases where at least one of two philosophical arguments must be wrong. In Chapter Three, I take on the widely cited argument that the dearth of female philosophers can, at least in part, be attributed to the fact that “women students are more likely than men students to find that their intuitions about the thought experiments discussed in their philosophy classes are at odds with those of their instructor” (Buckwalter and Stich, 2013).

1.3.3. Chapter Four: Factless, faulty disputes

Sometimes, what appear to be two contradictory views about a theory turn out to be two compatible but incommensurate views about two slightly different theories. In Chapter Four (Holtzman, under review A), I argue that previous characterizations of faulty reasoning in compatibilist belief may apply in some cases, but not in all of the cases in which philosophers seem to think that they do. Unless we know whether people take moral responsibility and free will to refer to different things in different contexts, we cannot tell the difference between a person who is irrational in the application of one concept, or rational in the application of two concepts for which the same word is used. And in many other cases, there are potentially shared facts about which disputants could directly disagree, but interlocutors are often discussing identifiably different facts without realizing it. When this occurs alongside the absence of a univocal rational structure for evaluating these facts, we have relativism, as mentioned earlier.
But this can also occur in the presence of a rational decision structure, as I argue is often the case when disputants fail to converge on a conclusive attribution of agency.

1.3.4. Chapter Five: Factive, faultless disputes

In Chapter Five, I then look at a case in which there seem to be objective facts of the matter, yet knowing the truth about those objective facts might do nothing to resolve certain disputes. This, I think, may be common when two sides are interpolating the normative-ethical implications of some descriptive fact. In Chapter Five specifically, it can be seen that even if we knew whether a certain gene did or did not increase the risk of recidivism, our factual discovery provides us with reasons that could reasonably lead to contrary courses of action. If a certain gene makes criminals more likely to recommit, we might infer that he should be kept behind bars to minimize harm to society; yet we also might infer that he should be left free since he was predisposed to reoffend by factors outside of his control. Here, the facts may be decisive but the reasoning is not; this is a case of monist relativism.

I devote Chapter Six to summarizing the taxonomy just laid out, and to setting for a plan of action for moving forward. Thus, my approach seeks to foster philosophical understanding by identifying the underlying causes of philosophical disagreement, in terms of component processes that are themselves philosophical disagreements.
When an *intuition* conflicts with an *analysis* of some concept, phenomenon, or course of events, is it always appropriate to say that we have *evidence* for an *objection* to that analysis? Analytic philosophers have, historically, taken the answer to be yes, but a number of recent works have cast doubt on this assumption. This chapter questions that fundamental assumption of analytic philosophy in a unique way, by identifying several ways in which personality informs philosophical belief. In the present study, individuals holding doctorates in philosophy were given a personality inventory and asked to respond to nine philosophical questions, seven of which produced significant sample sizes. Personality predicted response to three of these seven questions, suggesting that philosophers’ beliefs are determined in part by their personalities.

In first publishing these results, I took them to show that “at least some part of philosophy, however small, is subjective” (emphases added; Holtzman, 2013). This, roughly speaking, characterizes the present chapter as one in the vein of experimental philosophy’s negative project, the purpose of which is primarily to undermine the methods, claims to knowledge, and purported expertise of professional philosophers (Knobe and Nichols, 2013). I stand by my initial claim that part of philosophy is *subjective*, a term I use in here in an approximate sense, to refer to claims that, more strictly speaking, may be situated in disputes that
are simultaneously faultless and factless. Because I stand by my initial claim, I endorse reading this chapter as one with a largely negative thesis. But my negative claims are part of a broader, positive project: I also stand by my claim that some part of philosophy, however small exemplifies this characteristic subjectivity. The expansion of this initial research program in later chapters thus reveals that both the negative and positive projects emerge naturally from the study described in this chapter. While I have not, strictly speaking, revised the view or abandoned the thesis defended in this initial project, in each chapter I broaden my thesis to include everything discussed in the last chapter and, as such, narrow the range of philosophical questions to which I apply the term ‘subjective.’ My argument in this chapter, that we can accurately describe some philosophy as subjective, is meant to be quantified existentially, not universally.

In relation to the chapters that come after it, the main development of Chapter Two is the identification of individual difference variables that characterize groups of persons along dimensions that are value-neutral and non-essential to those persons, but which correlate with certain philosophical views. In later chapters, I plan to show how individual differences can be treated as surrogate variables, predictors that allow us to compare groups of people who, on average, can be expected to differ in their views about some philosophical question. Unlike in traditional psychology, the soundness of the initial argument for why some surrogate variable should predict philosophical judgment is not, strictly speaking, essential to interpreting the data, nor to making progress in understanding the precise issues raised by certain philosophical disputes. Such explanations can be psychologically valuable, and may help advance philosophy, but are not necessarily intrinsically philosophically interesting. However, the interaction of these variables with aspects of different questions asked—-in other words, the conceptual fault lines
within questions that lead different people to respond one way or another—are philosophically interesting.

Still, many skeptics about experimental philosophy believe that nothing culled from the intuitions of people not trained in philosophy should be expected to have any direct bearing on deep philosophical questions. Because these sorts of skeptics believe that the contributions of experimental approaches drop off precisely at the philosophical junctions they claim to address, I call this flavor of skepticism the deflationist approach. One advantage of the work reported in this chapter is that, because the participants all held PhDs or DPhils in philosophy, these standard deflationist responses become self-undermining—after all, if philosophers’ viewpoints in general don’t matter, then their views on deflationism don’t either. I will reference general perspectives on experimental philosophy from this approach later in this chapter in a bit more detail in §2.1.1. (Kauppinen, 2007; Devitt, 2011). Chapter Four describes a more detailed, specific kind of deflationism about empirical studies of the belief that causal determinism is compatible with free will is, a view known as compatibilism.

Traditional hypothesis-driven research and data mining can inform fruitful philosophical projects, so long as in both cases, sufficient replicability is demonstrated to assuage any potential methodological concerns associated with data mining (multiple statistical comparisons). Rather than concluding anything about personality per se, the purpose of this chapter is to develop hypothetical tools that may allow us to show that this alternate dimensions of the philosophical concepts in dispute. Here, I am only developing a tool to get at the core concepts, just like one might build a ladder to get at something otherwise out of reach. Later—at the ends of Chapters Four and Five—we will be able to “throw away the ladder” (Wittgenstein, 1921/2004).
2.1. The Nature of Philosophical Debate

Why do arguments that seem so plausible to some philosophers always fail to persuade others? Philosophical inquiry is supposed to consist in the rational pursuit of objective truth, so it is curious that even highly trained philosophers, versed in all the same arguments and thought experiments, disagree about so many points. To the extent that philosophical conclusions are reached through intuition and reason, the persistence of philosophical disagreement suggests that something very personal drives philosophical intuition.

In this chapter, I consider the view that philosophers’ beliefs are partially founded on the basis of individual character and disposition rather than objective evidence and irrefutable reason. The present study examined the relationship between philosophers’ personalities and their responses to several thought experiments. The data generated provide initial evidence that philosophers’ personalities inform their beliefs regarding specific philosophical problems, one of which is studied more closely in every chapter of this project. I consider several alternative interpretations of my findings, but ultimately conclude that personality sometimes affects

---

1 In order to avoid confusion, I will define my usage of certain key terms, consistent with familiar usage as indicated by the Oxford English Dictionary.

2 Rational: Based on universal norms of good reasoning, as opposed to emotion or individual preferences.

3 Personality: The set of qualities that distinguish an individual. These include, but are not limited to, stable, long-lasting, internalized traits. I will use the term ‘personality’ consistent with the broader set of personality traits; context will make clear when it is also consistent with the aforementioned subset.
philosophical belief. I further argue that some—though not all—differences in belief owing to personality and individual differences are rightly characterized as subjective. Insofar as this is true, belief about some philosophical problems is subjective, and insofar as belief regarding philosophical problems is central to philosophical practice, philosophy that relies on these intuitions can properly be called subjective.

2.1.1. Experimental Philosophy

Historically, philosophers have drawn conclusions about the world on the basis of intuition and formal reasoning. More recently, experimental philosophers have begun to apply scientific methodologies to philosophical questions in order to understand how people think about them. One problem with traditional philosophical approaches is that a particular philosopher may have peculiar and unpopular intuitions. In an effort to address this concern, experimental philosophers have recently begun polling average people in order to discover “folk intuitions,” a term that many experimental philosophers use to refer to the philosophical beliefs of the average person. What these experimental philosophers have found has led them to suggest a fragmented and manipulable “folk” (Cokely & Feltz, 2009a). For this reason, it is unclear to what the ‘folk’ in

---

4 Philosophical belief: Attitude towards philosophical problems. Response to such problems presented as vignettes is sometimes referred to as philosophical intuition, but this latter term carries a connotation of unreasoned immediacy, which is inappropriate to describe the response of philosophers to questions they have likely considered at length.

5 Subjective: Reflective of an individual non-privileged perspective, as opposed to entirely mind-independent, universal standards.
“folk intuitions” is supposed to refer to Cappelen (2013). I therefore use the phrase hesitantly, and only to refer to the work of others who use the phrase.

“Folk intuitions” have been found to vary across cultures (Machery, Mallon, Nichols, & Stich, 2004), to demonstrate frame effects (Nichols & Knobe, 2007), and to differ from the intuitions of professional philosophers (Machery & Systma, 2010). This chapter focuses on the relationship between personality and philosophical belief among philosophers, a relationship that has already been partially explored among lay people by Feltz and Cokely (2009a). They found that extraverts are uniquely predisposed to consider certain socially deviant acts harmful and immoral. They also identified a relationship between the personality trait openness and the belief that moral and physical facts may be subjective (Cokely & Feltz, 2008).

Because extraverts are thought to devote excessive attention to social aspects of situations, Feltz and Cokely also hypothesized that extraverts might lose sight of deterministic aspects of such situations (2009b). This, they argued, would cause extraverts to be particularly likely to consider free will and responsibility compatible with determinism. The authors asked undergraduates to consider a deterministic universe, and found that extraverts were indeed more likely than introverts to consider murderers free and to hold them responsible for their actions. Studies have since confirmed that this effect only occurs when participants are given high-affect, concrete scenarios, a finding that strengthens the claim that extraverts’ attention to social details is responsible for the effect (Feltz & Cokely, 2009b; Nadelhoffer, Kvaran, & Nahmias, 2009).

Unfortunately, findings like these have often been dismissed as trivial on the grounds that unlike lay people, philosophers are uniquely capable of reasoning in a formal, objective manner, or that they possess the requisite expertise required to answer philosophical questions (Kauppinen, 2007; Devitt, 2011). This study undermines those claims, and allays concerns that
the effects of personality on philosophical belief are circumscribed to social and moral questions, as none of the three effects identified in this chapter falls within those domains. Whether factors that increase people’s tendencies to judge free will and determinism compatible do so by supporting their core competencies, or instead by causing them to commit performance errors, is also major point of contention among experimental philosophers. These concerns cannot be addressed in this chapter, but they will be revisited in §4.4.

2.1.2. The Big Five Personality Traits

In later chapters, I will discuss arational factors (which may serve as what I have called surrogate variables) other than personality, such as gender and situation, as they relate to and apparently influence philosophical belief. For now, however, I confine my discussion to individual differences in personality. Allport and Odbert launched modern personality research by examining the English lexicon, and identifying 18,000 words that describe human behavior (1936). They chose to focus their research on the 4,500 or so words that describe stable, long-term personality traits, the kind that we often consider part of a person’s identity. Over several decades, they and others constructed numerous personality assessments, slowly eliminating highly correlated terms.

Tupes and Christal (1961) were the first to identify a five-factor model of personality, and by the 1990s a consensus was reached about the comprehensiveness of five dimensions: Extraversion (or Surgency), Agreeableness, Conscientiousness, Neuroticism (as opposed to Emotional Stability), and Openness to experience (or Intellect). Because of the enormous breadth of each category, the factors became known as the Big Five personality traits (Digman, 1990;
Further research has shown many of these characteristics to vary across cultures (Eap, DeGarmo, Kawakami, Hara, Hall, & Teten, 2008), between genders, and with age (John, Gosling, & Potter, 2003).

The precise definition of each trait remains controversial. Agreeableness measures several tendencies, including compliance, altruism, and supportiveness. Extraversion gauges how shy or outgoing a person is, and encompasses warmth, dominance, and sociability. There is debate as to whether conscientiousness is best understood as a sort of governor on impulsive behavior, or as an organizer of deliberate action. Neuroticism measures the frequency and extent to which individuals experience negative emotions, and the form and severity of their responses to these emotions. Openness measures intellect, aesthetic sensibility, creativity, and a number of other highly correlated factors (McCrae & John, 1992).

Several instruments have been developed to assess the Big Five, each with its pros and cons. The most commonly used assessment contains 100 unipolar trait descriptive adjectives; the most effective test, the Revised NEO Personality Inventory, contains 240 items. Recruiting participants to answer 240 questions online, and expecting them to do so with accuracy and attention, raises practical concerns, and so participants in my study were given the Big Five Personality Inventory (BFI). The BFI is a well-regarded personality survey composed of 44 questions (John & Srivastava, 1999).

### 2.2. Method

1,195 participants were recruited through the social networking site Facebook and the general interest philosophy blog Leiter Reports. Of those who completed the entire survey, this chapter
looks only at the 234 philosophers who held PhDs or DPhils in philosophy. Philosophers who participated were predominantly white (94%), male (82%), from Western or Australasian countries (100%), and were of all ages and socioeconomic backgrounds.

Participants were directed to a five-page survey on the website SurveyMonkey, where they were required to mark all questions but allowed to indicate a preference not to respond. On the first page of the survey, participants were briefed and asked to acknowledge consent and age of majority. The second page consisted of nine randomized “Yes” or “No” philosophical prompts, listed in the Appendix. Questions 5, 6, 7, and 9 borrowed heavily from famous thought experiments devised by other philosophers (Kripke, 1980; Gettier, 1963; Jackson, 1986; Thomson, 1976). Page three had 51 personality questions that were scored on a Likert scale from 1 (Disagree strongly) to 5 (Agree strongly), the first 44 of which comprised the BFI. These questions were not randomized, and BFI items were presented first and in the same order described in John and Srivastava (1999). Page four consisted of 7 demographic questions, and the last page thanked and debriefed participants. Because all responses were self-reported and unmonitored, there was the usual risk of false reports.

2.3. Results

Participants’ BFI scores were calculated from raw data as the average of responses given for each facet. Respondents who chose not to answer a given philosophical question were excluded from that question’s analysis. For each philosophical question, a binary logistic regression was run, in which all five personality traits and a constant were entered into the model. A summary of results for each overall model and each trait-belief pair is presented below (Table 1).
Table 1

Correlations Between Personality Factors and Philosophical Beliefs

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>A</th>
<th>C</th>
<th>N</th>
<th>O</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Compatibilism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>***</td>
</tr>
<tr>
<td>2. Fairness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Somatic Identity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Reductionist AI</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>***</td>
</tr>
<tr>
<td>5. Descriptivism</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Gettier/Knowledge</td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td></td>
<td>***</td>
</tr>
<tr>
<td>7. Knowledge Argument</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Embodied Cognition</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>*</td>
<td>**</td>
</tr>
<tr>
<td>9. Trolley Problem</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism; O = Openness. Numbered philosophical beliefs were tested with the corresponding prompts listed in the Appendix.

* $p < .05$

** $p < .01$

*** Bonferroni-corrected $p < .006$

In order to reduce the risk of false positives, each personality trait was force-entered into the regression, rather than entered stepwise. Only those questions for which a significant overall effect was found are considered below, in order to keep the number of initial comparisons to 9,
rather than 45. A Bonferroni correction for the use of 9 comparisons provided an even more conservative test. Question 3 (which did not exhibit significant overall effects) and Question 6 (which did exhibit significant overall effects) were excluded from further discussion because 10 or fewer participants responded “Yes,” to these questions, reducing the number of questions discussed from nine to seven.

For each regression model, $\chi^2$ measures the extent to which the observed pattern of responses differed from what one would expect if personality and philosophical belief were entirely independent. The associated Hosmer and Lemeshow $\chi^2$ reflects the extent to which the observed results differed from those predicted by the model. Greater $\chi^2$ scores reflect greater dependence of philosophical belief on personality, and greater Hosmer and Lemeshow $\chi^2$ scores reflect greater independence. The p-value of each of these scores represents the probability that the observed score was due to chance.

2.3.1. Dualism: Could a Robot Feel Love?

Participants were asked the following question:

Suppose neuroscientists are able to identify every part and every connection in the human brain. Working with a team of computer scientists, they then build a robot that has a complete electronic replica of the human brain. Could this robot experience love?

Most philosophers (73%, $N = 202$) believed that a robot with a replica human brain could feel love. The Big Five had a significant overall effect on this belief, $\chi^2 = 16.498, p < .01$, and
the model was a somewhat good fit, Hosmer and Lemeshow $\chi^2 = 14.522, p = .069$. As shown in Table 2, neuroticism and conscientiousness were both significant predictors of response.

**Table 2**

*Could a Robot Feel Love?*

<table>
<thead>
<tr>
<th>Personality Trait</th>
<th>$B$</th>
<th>$SE$</th>
<th>Lower</th>
<th>$Exp(B)$</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conscientiousness</td>
<td>-.916*</td>
<td>.37722</td>
<td>.191</td>
<td>.400</td>
<td>.839</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>-.728*</td>
<td>.294</td>
<td>.271</td>
<td>.483</td>
<td>.860</td>
</tr>
</tbody>
</table>

*Note. $B$ = Regression coefficient. $Exp(B)$ = Odds ratio.*

* $p < .05

$B$ is the value used to predict philosophical response from each significant personality factor. Negative values indicate that a factor is associated with disagreement, and positive values signal agreement. $B$ was also converted into the odds ratio, $Exp(B)$, which indicates that for each unit increase in a given personality factor (on a four-unit scale), philosophers’ odds of holding the target belief changed by that factor. For each unit increase in neuroticism or conscientiousness, philosophers’ odds of believing that a robot could feel love fell by more than half.

2.3.2. Descriptivism: Are Names Identical With Descriptions?

The next question on the survey asked the following:
Suppose that all you know about Einstein is that he developed the Theory of Relativity.

But suppose it turns out that Einstein actually stole the idea from some guy named Moynahan, who nobody has ever heard of. In this case, when you use the name “Einstein,” are you actually referring to Moynahan?

The predominant view among philosophers (87%, \(N = 202\)) was that the name “Einstein” could not refer to someone other than Einstein. Overall personality predicted response to this question, \(\chi^2 = 14.140, p < .05\), and the model was a very good fit, Hosmer and Lemeshow \(\chi^2 = 8.359, p = .399\). Agreeable philosophers were less likely to identify the name “Einstein” with Moynahan, the man who fit the description, but conscientious philosophers were more willing to assign Moynahan the name “Einstein” (Table 3).

### Table 3

*Are Names Identical With Descriptions?*

<table>
<thead>
<tr>
<th>Personality Trait</th>
<th>(B)</th>
<th>(SE\ B)</th>
<th>Lower</th>
<th>(Exp(B))</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness</td>
<td>-.922*</td>
<td>.364</td>
<td>.195</td>
<td>.398</td>
<td>.812</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>1.043*</td>
<td>.486</td>
<td>1.095</td>
<td>2.838</td>
<td>7.354</td>
</tr>
</tbody>
</table>

*Note. \(B\) = Regression coefficient. \(Exp(B)\) = Odds ratio.*

* \(p < .05\)
2.3.3. Embodiment: Could a Brain Think Without Being Connected to a Body?

The following question probed belief about the embodied mind thesis:

Suppose scientists are able to use stem cells to grow lungs that breathe without being connected to a body. They then grow a heart that pumps without being connected to a body. If they can do all this, can they create a brain that thinks without being connected to a body?

About half of philosophers believed that a disembodied brain could not think (54%, $N = 205$). Differences in overall personality modeled differences in belief, $\chi^2 = 15.375$, $p < .01$, and the model was a good fit, Hosmer and Lemeshow $\chi^2 = 3.220$, $p = .920$. Increases in agreeableness and openness were associated with more negative responses, indicating an increased belief in the embodied mind thesis (Table 4).

**Table 4**

*Could a Brain Think Without Being Connected to a Body?*

<table>
<thead>
<tr>
<th>Personality Trait</th>
<th>$B$</th>
<th>$SE B$</th>
<th>Lower</th>
<th>Exp($B$)</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness*</td>
<td>-.549*</td>
<td>.273</td>
<td>.338</td>
<td>.577</td>
<td>.985</td>
</tr>
<tr>
<td>Openness*</td>
<td>-.752*</td>
<td>.308</td>
<td>.258</td>
<td>.471</td>
<td>.861</td>
</tr>
</tbody>
</table>

Note. $B =$ Regression coefficient. Exp($B$) = Odds ratio.

* $p < .05$
2.4. Discussion: Do Personality Effects Mean Philosophy is Intrinsically Subjective?

The present analysis found that of the seven thought experiments considered, three were subject to overall personality effects. Running seven tests on a single data set can significantly increase the risk of a Type 1 error—a false positive—but the likelihood of finding two effects in seven tests, each significant to $\alpha = .015$, is itself significant to $p = .004$; the likelihood of finding three such effects is $p = .0001$. In a separate analysis presented at the 2013 Social for Personality Social Psychology Annual Meeting (Holtzman, 2013), a fourth thought experiment in this study showed specific hypothesized personality and gender effects. A fifth question has already been found by other authors to be subject to frame effects (Nichols & Knobe, 2007), in addition to being replicated, reported, and modeled in Chapter Four (Holtzman, under review A). In essence, philosophically irrelevant factors have been shown to affect intuitions in five of the seven cases tested here. It remains an empirical question whether the two remaining vignettes appeal to subjective factors, but given the countless number of psychological and sociological factors in play, it seems highly likely.

Regardless, the extent of the demonstrated effects shows just how dependent philosophical belief is upon philosophically irrelevant factors. Taken alongside findings that cultural upbringing and gender affect the philosophical intuitions of lay people, a picture emerges in which philosophical belief is subject to any number of personality factors. Participants in this study had devoted anywhere from five years to decades of their lives to studying philosophy at the highest level, but even this extent of philosophical training did not
eliminate certain arational\textsuperscript{6} prejudices.\textsuperscript{7}

The mechanism by which personality affects response to thought experiments deserves attention, but will not be discussed here. In what follows, I argue that because personality affects philosophical belief, philosophical belief is subjective. I then develop the further view that large swaths of philosophy are therefore also subjective. Finally, I consider four major but not exhaustive objections to the contention that personality affects philosophical belief, and conclude that none holds up to scrutiny.

\textit{2.4.1. The Connection Between Personality Traits and Subjectivity}

Beliefs affected by personality are by their very nature subjective. The fact that personality varies from person to person, and that there is no “right” personality to have, are central to the notion of personality. If philosophical beliefs are adopted in part on the basis of their appeal to non-objective values or motivations such as those measured by the BFI, then it is only fair to describe these beliefs as partially subjective.

Consider the possibility that personality traits affect philosophical beliefs without lending them any sense of subjectivity—that philosophical problems are simply underdetermined, not

\textsuperscript{6} Arational: Not within the domain of rational analysis, and therefore neither rational nor irrational. The results of arational processes may accidentally appeal, or fail to appeal, to rationality, but that this is a separate matter.

\textsuperscript{7} I take “prejudice” to be capable of referring to both favorable and unfavorable inclinations. Prejudice, as understood here, may lead its subject towards or away from the truth, or may apply in the absence of any objective standards at all.
subjective. A doctor, for example, may maintain objectivity when choosing a course of treatment even if she is unsure what will work best. She has to fill in the unknowns somehow, but every doctor does, and this does not give us reason to call doctors’ opinions subjective. If philosophical problems are likewise underdetermined, a philosopher could fill in the blanks on her own without properly being described as subjective. Unfortunately, this argument is convincing only in its vagueness.

My findings demonstrate that particular personality traits predict particular philosophical opinions. Therefore, a better analogy would be the discovery that the more neurotic a doctor is, the more likely she is to recommend prophylactic surgery, or that the more agreeable she is, the more likely she is to dispense painkillers to anyone who asks. These are obvious medical biases, ones so glaring and deleterious to medicine that they could present serious ethical concerns. To deny that my findings are equally worrying for philosophy is wishful thinking.

2.4.2. The Connection Between Philosophical Belief and Philosophy

It certainly seems as though many important zeitgeist shifts in 20th century philosophy owe primarily to thought experiments and the specific beliefs they elicit. The purpose of Kripke’s Gödel cases was to disabuse the descriptivist masses of beliefs Kripke thought to be mistaken. Jackson’s black-and-white room might not confine anyone to a unique view, but it would be difficult to argue that every belief about his thought experiment is consistent with every set of philosophical beliefs. Still, it is an empirical question whether deep philosophical theories demonstrate the same effects as philosophical vignettes, and whether vignettes actually affect philosophers’ views, and I can’t claim to answer those questions here.
Regardless, philosophy’s purpose is not merely to construct general theories about the structure of the world. It is also to describe the actual world in ways consistent with those theories. It is the rationality and objectivity of this latter aspect of philosophical practice—in which ethicists claim to know what is the right thing to do, and metaphysicians claim to know who is whom—that this study impugns. Insofar as philosophy purports to provide tools for understanding our world, the subjectivity of philosophical belief in a given domain can properly called the subjectivity of philosophy in that domain.

Of those philosophical endeavors that utilize intuitions about cases, some may be more prone to personality effects than others, and some personality effects may be more rightly characterized as evidence of the subjectivity of philosophy than others. The objective evidence in favor of some beliefs may be so strong as to eliminate most or all subjective effects, but many problems fall somewhere along a spectrum of subjectivity. Just how much of philosophy is subjective, and just how subjective, remains an open question.

One might protest that philosophers could be subjective without causing philosophy to be. It is possible, after all, for individual doctors to be subjective without endangering the objectivity of medicine. However, the same cannot be said of philosophers and philosophy. Medical treatments and experiments yield observable outcomes, which provide an objective indication of which notions are right and which are wrong. But for some philosophical problems, the only data we have are our beliefs. Without any impersonal check on personal views, the subjectivity of practicing philosophers can rightly be called the subjectivity of philosophy.
It is important to determine whether the personality differences identified here truly predict philosophical disagreement, or merely differences in gut reaction that lead to disputes between views that are, strictly speaking, incommensurate rather than contradictory or mutually exclusive. Only in the former case would these findings tell us something meaningful about philosophy. There are three distinct questions in this vein.

First, do neurotic philosophers object to certain ascriptions simply as a matter of antagonism? Jesse Prinz and Joshua Knobe (personal communication) have both suggested that neurotic philosophers might simply object to a robot feeling love out of resistance to liberal use of the term “love,” not necessarily as expressions of their deeper philosophical views regarding the relevant phenomena. If so, this might only be a case of philosopher bias, rather than philosophical subjectivity. However, this explanation cannot be right in this particular case. Antagonism is defined on the BFI as the opposite of agreeableness, and though neuroticism and agreeableness have been shown to correlate slightly negatively (-.28) (John & Srivastava, 1999), agreeableness itself did not predict response to this particular question.

Second, when disagreeable philosophers respond ‘No’ to questions, are they expressing genuine theoretical beliefs, or merely being negative? Disagreeableness actually predicted a ‘Yes’ response to both questions for which agreeableness was a predictor of response, so this worry also seems unfounded.

Finally, when disagreeable philosophers express dissident views, are they expressing genuine theoretical differences from the mainstream, or merely being curmudgeonly? In this study, antagonism correlated with heterodox responses to both questions for which it was a
predictor. Whether or not this is because disagreeable philosophers are preconditioned to adopt unpopular views is unclear, but even if that were the case, that would only bolster the claim that subjective factors influence professional philosophers’ philosophical beliefs.

While this shows that many personality effects on philosophical belief may be of philosophical import, this does not mean that that all such effects are. Cases of subjectivity, in which the prejudices of different lived experience are inseparable from the philosophical questions themselves, are just one kind of case. There are surely cases of bias in which an objectively false view is embraced by some theorists for reasons that are not only irrelevant to philosophical inquiry, but which turn out to undermine the pursuit of truth.

For example, it would not be surprising if, prior to the Copernican revolution (and perhaps also after it), belief in a geocentric universe was highly correlated with religiosity. The idea of angels pushing with crystalline spheres in the heavens above a centralized Earth has more than a modicum of theistic appeal. But any potential correlation between geocentrism and religiosity does not render the former subjective, nor its disputes factless. Facts pertaining to the truth or falsity of the relevant astronomical propositions theoretically can be (and, arguably, have been) established independent of individual differences in religious experience, faith, and practice.

In cases like this, personality effects (should any exist) clearly must lead toward or away from the singular, identifiable, objective falsity of certain geocentric claims about epicycles. There can be no biasing toward or away from heliocentrism that is not, properly speaking, likewise a biasing toward or away from truth. Apparent retrograde motion simply does not occur because we observe, from our own stationary planet, other planets moving along spheres within spheres. Retrograde motion occurs because both our planet and other planets are racing in
ellipses around the sun. We occasionally “lap” planets more distal from the sun than we are, causing them to fade backwards in appearance, like cars being passed on the highway. Facts like these are true or false independent of subjective experience.

2.4.4. Objection: The Big Five Do Not Truly Assess “Personality”

Neuroticism, for instance, is not wholly reducible to the kinds of questions asked on personality inventories, and a highly neurotic person could score lower on this facet of the BFI than someone we would not consider highly neurotic. Though the BFI measures certain prominent personality traits effectively, one might hesitate to grant that it measures extraversion, agreeableness, conscientiousness, neuroticism, and openness per se. And even if the BFI were a perfect metric, the kinds of self-reports used here face obvious limitations.

Nonetheless, the objection is myopic. Even if BFI scores do not perfectly reflect the personality traits they claim to, they still evaluate important personal inclinations. And even though self-reports may be biased, they still reflect the personality of the subject. In fact, even if the BFI reflected nothing that it purports to, its correlation with philosophical beliefs would still be disturbing. BFI scores reflect highly reliable subjective differences, and these subjective differences predict philosophical differences. Supposing that all of the above accusations are true, this study would still reveal a meaningful relationship between subjective factors and philosophical beliefs.

---

8 I intend ‘biased’ to be interpreted without its occasional connotation of being skewed towards or away from the truth. I simply mean skewed, regardless of any objective standards.

9 For definition and estimates of reliability, see John & Srivastava (1999).
2.4.5. Objection: Personality Does Not Affect Philosophical Belief

It is possible that my theory reverses the order of causation, and that philosophical belief affects personality and not the other way around. However, this is highly unlikely for a number of reasons. BFI scores remain relatively stable throughout a person’s lifetime, not changing much even after education (John, Gosling, & Potter, 2003). Personality develops long before philosophical belief, and lifelong characteristics cannot be informed by factors that form later. And many people may never think deeply about philosophical problems, despite having rich personalities.

One could go a step further and suggest that personality and philosophy do not affect each other at all, and are instead shaped by some common force. But if such a force shapes personality, it could only be described as affecting a person’s subjective values. Appealing to such a force, then, does little to support the claim that philosophical values are objective.

2.4.6. Objection: The Philosophical Prompts Were Not Worded Fairly

The prompts used in this study were created to be as neutral as possible, but are not perfect. Some readers will undoubtedly decry that the questions that were based on famous examples seem contrived and unrealistic. Other questions may appear to bring in concerns beyond those they claim to test. I will not defend the prompts against these charges, because I think the charges are apt. I believe that many of the vignettes may have been loaded or unrealistic, or at least failed to control for philosophically irrelevant factors. I admit that they appealed to certain sensibilities that may have clouded the issues. How else could I have found an effect?
This is the nature of philosophical thought experiments. Intuition pumps often take the fore in philosophical debates, despite distorting the issues. The claim that even the purest philosophical cases bring philosophically irrelevant factors into play does not weaken my argument—it is central to my argument. As such, the objection that these vignettes evoked biased responses is not really an objection at all. Subsequent studies could vary the wording of the prompts, but it is not clear what this would accomplish. If the same results were observed, the findings of this chapter would be bolstered. But if the effects disappeared, that would only suggest that framing and personality interact in a profound way to influence philosophical belief.

Such interactions between seemingly irrelevant facts and individual difference factors are, in fact, central to the analyses discussed in Chapters Four and Five. Researchers who are interested in positive experimental philosophy projects should actively seek out these kinds of interactions. For example, one form of conceptual analysis, adapted from Carnap’s scientific notion of *explication* (1950), seeks to refine ordinary language philosophical concepts into more precise ones that “retain similarity of conceptual content with the explicandum, and increase precision, fruitfulness, and simplicity” (Shepherd and Justus, 2014). Refining ordinary language concepts into new concepts that maintain similarity to the old ones, that are precise in distinguishing the views of disputants, that are fruitful in resolving those disputes, and that simplify philosophical problems are all ends that the project presented here can help us achieve.

One example of an explicatory interaction will be discussed in Chapter Four (Holtzman, under review B). There, it is revealed that two descriptions of a person’s causal history tend to be treated as equivalently exculpatory by one group of people, but differentially exculpatory by another group. This raises important questions about how the different lived experiences of these two groups may have given rise to these different tendencies in philosophical belief. This has
implications, for example, for all four goals of *explication*. Retaining similarity of conceptual content will require us to identify and retain only that content that tends to overlap between our two groups’ concepts. Increasing precision will require us to identify why and over what principles their two different readings emerge and fail to overlap. Increasing fruitfulness will mean finding a way to establish conditions such that our two groups, when employing our two descriptions, are less frequently at cross-purposes and simply talking past one another (Kuhn, 1970). And simplicity, quite simply, may have to be sacrificed if we are to attain our first three goals.

### 2.4.7. Conclusion: Some Parts of Philosophy Are Subjective

Thought experiments are designed to be the great deciders in many philosophical debates. When clever philosophers are convinced of a certain view, they often design thought experiments to demonstrate just how intuitive their positions are. But other philosophers invariably disagree about the results of those experiments, often for reasons they cannot explain. At such pivotal moments, their minds are already made up, in part by their unique personalities. This is why views that seem obviously right to some philosophers seem obviously wrong to others. Despite years of training, philosophers see certain problems from personal perspectives. To the extent that these philosophical stalemates are the result of intrinsic human limitations, they may never be resolved.

It has not been my intention here to claim that there are no objective philosophical truths, nor to claim that there are. I have also not attempted to identify the locus or mechanism of philosophical subjectivity, although I think that these are rather interesting problems. I only hope
to have argued at least some part of philosophy, however small, is subjective, that personality
effects are one manifestation of this subjectivity, and that philosophers may need to reappraise
the ability of philosophy to discover objective truths.

Philosophy’s aim has traditionally been to discover truth through reason and intuition
alone. I have shown that reason and intuition are intimately tied to personality. It remains to be
seen whether reason and intuition, philosophical or otherwise, can be divorced from personality
factors, and I have only shown relationships in a limited number of cases. Still, the burden of
proof now lies with the defender of traditional philosophical methods. In order to defend
philosophy, philosophers must first defend the autonomy and transformative power of reason.

2.5. Implication: Some Parts of Philosophy Are Not Subjective

Earlier, I expressed my central thesis in terms of four kinds of philosophical disputes, noting that
subjective disputes (which are both factless and faultless in nature) constitute just one of these
kinds. Still, these results pave the way to study the other three kinds of dispute outlined earlier.
In some cases, we will be fortunate enough to be able to construct a potentially sound and valid
argument by adding a single, empirically testable premise, and then demonstrating its falsehood.
Such is the case in the next chapter.
Lately, there has been a good deal of discussion about why there are so few women in professional philosophy. In particular, special attention has been given to a hypothesis put forward by Buckwalter and Stich (2013), who suggest that undergraduate women tend to have different philosophical intuitions than their professors, and that this leads them to leave the field. My first aim in this chapter is to debunk that idea. My second aim is to put forward an alternative hypothesis.

3.1. Rejecting Beliefs, or Rejecting Believers?

It has been discovered that philosophy PhDs are more likely to reject certain rule violations if they are especially neurotic (Holtzman, 2013). It has also been found that the opinions philosophy faculty express toward certain dilemmas can easily be swayed by presenting them with other, similar dilemmas first (Schwitzgebel and Cushman, 2013; Schwitzgebel and Cushman, in process). Why, then, should we expect that bias against an interlocutor, a student, a classmate, or oneself would have any less sway on the opinions we come to form about the philosophical views expressed by others? My hypothesis is that we should not expect that at all.
The idea that disposition plays a role in the establishment of philosophical beliefs has frequently been misinterpreted—often uncharitably—by both detractors and supporters of experimental philosophy. For detractors, caricature interpretations and straw man arguments have played a crucial role in criticisms of the movement. Those who inveigh against this view seem especially drawn to a line of attack that we might call the *insufficiency of sentiment* argument, which has deep roots in modern philosophy:

> It is often necessary, we find, that much reasoning should precede, that nice distinctions should be made, just conclusions drawn, distant comparisons made complicated relations examined, and general facts fixed and ascertained (Hume, 1777/1983).

But the insufficiency of sentiment is a point that has been elucidated and endorsed by even the staunchest sentimentalists—in fact, Hume was the archetypal sentimentalist. So this worry can do little damage against that view when wielded against critics of experimental philosophy.

Experimental philosophers themselves have often overlooked the distinction between *fundamental faultlessness*, and *total incorrigibility*. Now, say a philosophical dispute is fundamentally faultless if it is absolutely resistant to adjudication by appeal to objective reason. This definition is qualitative—it says only that there is something in the fundament of a disagreement against which appeal to rationality is essentially futile. But this in no way implies that either of the disagreeing views is totally, or even mostly, incorrigible. Notably, faultlessness concerns incommensurate sets of utterances that, when contrasted with one another, give rise to almost-but-not-technically arguments, and which may include utterances that are almost-but-not-quite propositions. In contrast, incorrigibility refers to attitudes toward sets of utterances—
attitudes that, if incorrigible because they are rooted in arational processes that are not truth-evaluable, cannot rightly be called beliefs. This leads us to an important contrast I wish to draw between a set of views that I endorse, and set of views that I reject. I endorse a thesis of faultlessness in many areas of philosophy, according to which differences between philosophical theories are essentially unadjudicatable. I reject the incorrigibility thesis, which I take to be a view about persons, according to which the conflicting theoretic endorsements people make are essentially unrevisable.

Hume recognized that for all the import of individual differences in determining moral character, education may play an even larger role. The juxtaposition of individual differences in disposition, and diversity of educational experience forms the crux of this chapter. Here, my intent is to demonstrate the theoretic myopia and practical problematicity of misinterpreting the fundamental faultlessness of theories as the incorrigibility of theorizers. I intend this exploration to double as a cautionary tale of what happens when one crosses the line from characterizing disputes as faultless to characterizing persons and their views as incorrigible.

3.2. Focus on the Expert/Gender Assumption

Why are there so few women in philosophy? Lately, there has been a good deal of discussion about the causes of this undesirable state of affairs. In particular, special attention has been given to a hypothesis put forward by Buckwalter and Stich (2014), whose view has been summed up beautifully in a takedown by Louise Antony:

Their idea is that if women have different intuitions about standard thought-experiments
than men do, and if men dominate philosophy, then women studying philosophy may come to the conclusion—or be told explicitly—that they just don’t “get” philosophy—that philosophy is not the subject for them (2013).

My central aim in this paper is to debunk the idea that women tend to have heterodox philosophical intuitions, and that these intuitions, “when combined with one of the standard methods invoked in doing and teaching philosophy, have the potential to generate unconscious and unintentional biases against women.” I think it much more likely that the nature of philosophical debate simply enables general, pre-existing biases against women to influence philosophy professors more strongly than they can influence professors of academic disciplines in which appeal to objective truths (of the kind “2 + 2 = 4”) is more straightforward. Crucially, this is not to say that the availability of objective truth is a panacea. Just as DNA evidence from thousands of crimes against women sits untouched on the shelves on evidence rooms across the country, objectively true answers classroom may be easily in the classroom ignored by professors and students with preexisting biases.

In §3.2., I draw attention to a misguided, unsubstantiated, and evidently false assumption which, from a logico-deductive standpoint, is required by the argument mounted by Buckwalter and Stich. Although those authors never explicitly acknowledge this premise, it is strictly entailed by their empirical claims. Yet as I argue in §3.3., the premise is fundamentally misguided inasmuch as it implicitly undermines the very egalitarian advocacy that their empirical work is intended to support. In §3.4., I show their assumption to be unwarranted in the sense that the evidence they present is compatible with two competing propositions, neither of which they acknowledge, let alone falsify, and both of which contradict the premise they assume.
I then reanalyze some of Buckwalter and Stich’s own data in §3.5., in order to show that the very evidence that they claim supports their view actually weighs against it. In §3.6., I suggest some lessons that the burgeoning field of experimental philosophy could learn from the mistakes made by Buckwalter and Stich.

3.2.1. Rehashing the argument

Adelberg et al. (2014) characterize the argument in Buckwalter and Stich (2014) as follows:

(1) If women have different intuitions about philosophical thought experiments than men, then this would likely lead more women than men to stop taking more philosophy classes.

(2) Women do have different intuitions about philosophical thought experiments than men.

(3) So, more women than men are likely to stop taking philosophy classes, which is one cause of the underrepresentation of women in philosophy (Adelberg et al., 2014).

Their project is meant as a theoretical refutation of premise (1) and an empirical refutation of premise (2), so their paper is doubly potent. Still, some readers might not be convinced by either of the arguments provided by Adelberg et al. However, there remains a third premise—the
expert/gender assumption—which is necessary to infer (3) on the basis of (1) and (2). I intend to show that this assumption holds up to neither logical nor empirical analysis:

When male and female undergraduates have divergent philosophical intuitions, women’s views differ from those of their professors to a greater extent than men’s views do.

I intended to argue against this premise, but first, I argue that the expert/gender principle assumes that gender differences remain constant across training belies an unwarranted essentialism, inasmuch as it assumes that male philosophical thinking is continuous across levels of training. Second, I show that whether one takes the expert/gender assumption to be essentialist, or merely one about intrinsic gender differences (as distinguished in Antony, 2013), its employment in advocating the greater representation of women in philosophy would appear to reduce that very argument to absurdity. Finally, I show that even if one does not accept either of these philosophical disputations, the conclusions of Buckwalter and Stich should still be rejected, because their critical assumption—that the evidence they provide supports the expert/gender principle—is directly undermined by some of the very data on which they build their case.

3.2.2. Revisiting the Expert/Gender Assumption

If most philosophers are men, then shouldn’t philosophers tend to have views like those of male undergraduates? This seems to be taken for granted by Buckwalter and Stich, and if true it would indeed justify the expert/gender assumption. But the assumption itself entails a troublesome
corollary. The expert/gender principle implies that as women develop philosophical expertise, they learn to think more and more like their male classmates already do. If this turned out not to be true, then there would be no reason to think that women’s intuitions were especially likely to be “unpopular” in the classroom in the way Buckwalter and Stich claim they are. Although Buckwalter and Stich would probably reject the claim that philosophers and male undergraduates tend to think alike, while female undergraduates are philosophical black sheep in need of a shepherd, this is what their view entails. Antony (2013) has already laid out an excellent and I think convincing critique of this “different voices” hypothesis, but even if the expert/gender assumption were warranted, it could only provide two reasons to encourage more women to enter philosophy. I do not think either would be embraced by Buckwalter and Stich, because their message is ultimately one intended to bolster equality.

First, if philosophical expertise is like mathematical expertise, in the sense of having intrinsic value and objective criteria of evaluation, then the expert/gender principle implies that men are “naturals” who require less training to become experts, whereas women can at best hope to catch up to men with training, but often find it too hard and quit. This reasoning seems both ill-founded and illogical. It is ill-founded because it would be inaccurate to describe the unpopularity of women’s answers in algebra class as the cause of their relative scarcity in upper-level college mathematics courses, and a similar explanation of the gender imbalance in philosophy seems equally implausible. And it is illogical because if women actually were intrinsically worse than men at math, it would be imprudent to cite their impoverished skill as a reason in favor of their greater inclusion in the field. So if philosophy is like math, Buckwalter and Stich are advocating the inclusion of more women in the field precisely because they are bad
at it, and this makes no sense. Fortunately, I think Buckwalter and Stich have made it clear that they do not think philosophical expertise is of a kind with mathematical expertise.

Second, the authors have elsewhere expressed the opinion that the intuitions of professional philosophers are no more reliable than those of anyone else (Buckwalter, Stich, and Tobia, 2013). The consideration that philosophical training is not like mathematics dissolves a substantial portion of the potentially normative implications of their “different voices” view, but it would also seem to render their normative argument for balancing gender in philosophy absurd. Efforts over the past few decades to integrate women into math classrooms have successfully reduced the gender gap in math performance, and these efforts were motivated by an understanding that women were being deprived, discriminated against, and provided with opportunities unequal to those provided to men. But if philosophy is less like math and more like, say, whittling, the only two reasons to advocate for more women in the field would be that men do it and so women should do it too, or that not enough women do it and that’s reason enough why more should. Neither of these is a satisfactory call to action for fixing what may be the result of inegalitarian treatment and complex social dynamics.

3.2.3. Reframing the Discussion

The assertion that we ought to correct the gender imbalance in philosophy is compatible with the view that feeling excluded, the cause put forth by Buckwalter and Stich, is not the issue. I have suggested that pandering to “women’s intuitions” is an uncalled for, and quite possibly incoherent proposition, but that does not mean that no action should be taken. The idea that feeling excluded is the reason more women choose to leave philosophy not reflects a somewhat
low opinion of women. Why presume that women leave because “they don’t get it,” rather than because they “already know that.” It also requires us to make general assumptions about the psychology of women. If nothing else, this seems like a much more difficult task than looking at the behavioral data in front of us.

Right now, you are looking at evidence that women are being excluded from consideration as expert philosophers. There is no need to formulate an opinion of women, or to guess at the psychology of others, to see this self-evident truth. A deeply and sometimes implicitly held prejudice against women-as-experts is not only an alternative to the expert/gender principle, but also a characterization of it. Buckwalter and Stich exclude the very possibility that women might perform adeptly in philosophy classrooms when they assume that:

(a) In developing philosophical expertise, men’s views tend to remain fixed whereas women’s views tend to become more like those of male undergraduates.

But this assumption is no more or less warranted than its contradiction:

(b) In developing philosophical expertise, women’s views tend to remain fixed whereas men’s views tend to become more like those of female undergraduates.

And both of these hypotheses are less parsimonious than a null hypothesis that, absent any evidence to the contrary, is by definition statistically more likely than the other two possibilities combined:
In developing philosophical expertise, people do not learn to think more and more like their classmates of any particular gender already do.

Buckwalter and Stich state that they see no rhyme or reason in the patterns of difference in their data, and so it may appear that they reject the notion of stereotypically feminine intuitions. But an unseen thread runs through all the cases for Buckwalter and Stich. Their entire argument is built on the assumption that women’s natural philosophical beliefs are, compared to those of their male peers, decidedly inexpert. Why not test their assumption against some of their own data?

3.2.4. Reanalyzing the data

Buckwalter and Stich claim that “there is little or no reliable data concerning professional philosophers’ intuitions in these cases” (2014), but this is false. Instead, Buckwalter and Stich seem to have fallen prey to a selection bias, having simply ignored data that were in their possession, and that undermine their view. This perhaps innocent failure to recognize the opportunity to falsify their hypothesis was compounded by the omission of data they collected but which failed to conform to their predictions. This latter methodological shortcoming represents a second, special kind of selection bias, which raises an issue known in the social sciences as ‘file-drawer problem.’ The most common way the file-drawer effect leads to false positives manifests is due to the hesitancy of journals and researchers to publish null effects (Rosenthal, 1979). But here, as Buckwalter and Stich seem to have overlooked data provided to them that undermined their a priori hypothesis.
Three of the nine externally discovered gender differences they report were drawn from a study that included over 200 participants who held either PhDs or DPhils in philosophy. Professional philosophers’ views on these cases have since been published (Holtzman, 2013), as have their attitudes toward six additional vignettes included in the same study. Buckwalter and Stich presumably analyzed all nine vignettes in that study, but chose not to report results from cases in which they were unable to find the undergraduate gender differences. Buckwalter and Stich also had access to data about the responses of male and female non-academics and philosophy professors to eight additional ethical dilemmas, but chose not to report these data, either. These data have been available online for several years, and reveal gender differences in only one of eight cases (vegetarianism); and in that one case, women were more likely than men to respond in the same way as professional philosophers (Schwitzgebel and Rust, 2011).

To see if the women discussed in Buckwalter and Stich’s paper actually thought less like professional philosophers than the men did, I reanalyzed the sole dataset cited in their paper for which this hypothesis was actually testable, as it was the only one that contained data from philosophy PhDs. Three of their nine external results were originally tested for “Do Personality Effects Mean Philosophy is Intrinsically Subjective?” (Holtzman, 2013), which measured the responses of philosophers and non-philosophers to nine “Yes” or “No” philosophical questions.

 Critics of Buckwalter and Stich have called for public access to the results they collected but did not include in their paper, and so I have include all nine results from Holtzman (2013) here, which reflect responses to the full questions that appear in that paper’s appendix. For all analyses below, chi-squares were conducted to compare the raw mean responses (Cohen, 1988) of male non-philosophers ($n = 104$), female non-philosophers ($n = 93$), and total philosophy PhDs who indicated gender ($n = 232; 17.3\%$ female). Pairwise chi-squares comparing scores of
non-philosopher men and/or women to overall PhD scores are reported for all cases in which those tests achieved statistical significance. For cases in which this did not occur, chi-square scores for the overall 3x2 table are presented. All charts represent the proportion of each population sample whose answer implied agreement with a particular view.

Omnibus tests showed that among non-philosophers who answered all nine questions, across the nine cases, women were significant more likely than men to share the opinions of professional philosophers $F_{22, 152} = 5.018, p < .001$. Of the three cases reported in Buckwalter and Stich, none supports the hypothesis that women tend to have different intuitions about thought experiments than professional philosophers do. When asked to make an abstract judgment of compatibilism, only the responses of men were significantly different from those of professional philosophers (Figure 1), and this same gender pattern arose for the knowledge argument (Jackson) (Figure 2). When asked to assess whether love could be reduced to a program a robot could run, the same overall trend appeared but no pairwise differences achieved significance (Figure 3).
Free will is compatible with determinism

- Male: 35.3%
- Female: 63.3%
- PhD: 55.6%

Figure 1. Column marked // was significantly different from PhD ($\chi^2_{315} = 11.405, p < .001$).

Taste qualia might be scientifically reducible

- Male: 39.2%
- Female: 17.2%
- PhD: 27.1%

Figure 2. Column marked /// was significantly different from PhD ($\chi^2_{319} = 4.802, p < .05$).
In three more cases from Holtzman (2013) not reported by Buckwalter and Stich, men but not women gave responses that were significantly different from those of philosophers. The first of these questions asked whether fairness in a certain case required adherence to the minimax principle (Figure 4), and the other asked whether it might be theoretically possible to create a brain that thinks without being connected to a body (Figure 5). In another case not reported by Buckwalter and Stich, in which subjects were asked to judge the ‘switch’ condition of the trolley problem (Holtzman, 2013), only women demonstrated a pattern of response significantly different from that seen among philosophers (Figure 6).
Minimax is no fairer than maximizing average utility

Figure 4. Column marked /// was significantly different from PhD ($\chi^2_{321} = 4.673, p < .05$).

An unembodied brain might be able to think

Figure 5. Column marked /// was significantly different from PhD ($\chi^2_{298} = 4.580, p < .05$).
Figure 6. Column marked /// was significantly different from PhD ($\chi^2_{284} = 6.935, p < .01$).

Two additional cases—a Gettier case regarding whether justified true belief might fall short of knowledge (Figure 7), and a Kripke case about whether names must be associated with descriptions (Figure 8)—elicited responses from both men and women that were significantly different from those given by professional philosophers. A final case, which contrasted the somatic and psychological conceptions of identity, revealed no significant difference in belief between any of the groups (Figure 9).
Figure 7. Columns marked /// were significantly different from PhD (Male $X^2_{321} = 4.265$, $p < .05$; Female $X^2_{313} = 3.902$, $p < .05$).

Figure 8. Columns marked /// were significantly different from PhD (Male $X^2_{328} = 5.666$, $p < .05$; Female $X^2_{319} = 19.573$, $p < .001$).
Overall, there were four cases in which men’s views differed significantly from those of philosophers but women’s views did not. There was also one case where women’s views diverged but men’s did not, and four cases in which neither or both groups had significantly different views from those of professional philosophers. What should we make of this dataset as a whole? At this point, the only unreasonable interpretation of these data would be certainty about any hypothesis. My purpose in presenting these data is only to refute the notion that we have any reason to accept the expert/gender assumption, not to draw statistical inferences from this dataset to any positive hypothesis.

Figure 9. Overall $X^2_{397} = .363, p = .834$.

3.3. Rethinking the Moral of the Story
My positive view, then, is this: Pursuing an understanding of what characterizes the philosophical psychology of different groups, and working to improve working conditions in academia, are both important projects. However, convoluted arguments should be carefully attended for inadvertent implications that might be both false and easily falsifiable. Such premature claims can only stymie further, more conscientious research. This is a general point for experimental philosophy. It is true that analytic philosophers have historically given voice to certain groups more so than others. But if we really want to make progress, and to eliminate bias from the field, more of the same will not do the trick. Instead of using data to defend armchair intuitions, we should let the data speak for themselves. Even though good researchers like Buckwalter and Stich are interested primarily in uncovering the truth, and do not set out to defend their own *a priori* intuitions, it is all too easy for confirmation bias to creep in unless we treat our own views as alternative hypotheses, to be considered plausible only once we have rejected the null hypothesis.

Readers who agree would do well to read Adelberg *et al.* (2014). In addition to raising awareness of the dangers inherent in using data mining techniques like those employed in Buckwalter and Stich, they discuss three carefully considered ‘Further Problems’ with the explanation given by Buckwalter and Stich. Adelberg *et al.* developed these three alternative views scientifically—in the absence of evidence for or against these empirical possibilities—and treat their adjudication as open questions. Their work provides an exemplar of how experimental philosophers might use armchair intuitions to motivate scientific explorations in a more impartial way than Buckwalter and Stich do. It is ultimately a good thing that Buckwalter and Stich have spurred this much-needed discussion about conditions for women in philosophy, and extending...
their project in the way Adelberg et al. have can only help improve our understanding of the situation.

3.4. Building on the Results

These results cannot tell us the causes of any potential gender differences in philosophical belief, and they are insufficient to direct us toward any particular course of effective action to remedy such problems. Nonetheless, they are valuable, and can serve as the basis for theories and research in any number of directions. It would be both interesting and useful to know if there is a pattern to these gender differences, and if so, how we might characterize that pattern of difference and its causes. This kind of project is a psychological one, and while it interests me greatly, my purpose in this project is to understand the nature of philosophical concepts and disputes. Thus, I will unfortunately have to turn away from more socially relevant questions (though I will return in Chapter Five), and turn toward more abstract, philosophical ones. What is it about the philosophical vignettes discussed in this chapter and the last that leave so much undetermined that facts about the persons considering those vignettes can consistently play a role in the way they assess those vignettes?

One possibility, as I have suggested, is chance, whose influence grows with the kinds of repeated observations used by Buckwalter and Stich. Ruling out chance is the first step toward developing an alternative theory. To mitigate the role of chance, it is important to narrow the scope of our questioning, and this is part of why I will focus more narrowly in the rest of this project on issues relating to determinism, free will, and moral responsibility. In particular, I
examine the factors that lead people either to embrace or to reject the philosophical thesis that causal determinism is compatible with free will and with moral responsibility.
With our pre-scientific concepts we are very much in the position of our archaeologist in regard to the ontological problem. We have, so to speak, forgotten what features in the world of experience caused us to frame those concepts, and we have great difficulty in calling to mind the world of experience without the spectacles of the old-established conceptual interpretation. There is the further difficulty that our language is compelled to work with words which are inseparably connected with those primitive concepts.

-Albert Einstein, “The Problem of Space, Ether, and the Field in Physics”

The positive project of experimental philosophy usually aims to replace the question of what people should believe with one about what people do believe and, in its most extreme form, to reduce the former to the latter. My proposal here has been to shift focus onto a question that I think is more fundamental than, and crucial to, all three of these projects: Why do people disagree about the answers to philosophical questions?

Many answers have been proposed, and my suggestion is that there in fact are many answers, depending on what kind of philosophical question is being asked. My proposal is to deemphasize the central question of what people should believe, and focus instead on why they believe it. To resolve disagreements, it might not be best to ask why we disagree about all philosophical problems, nor to only ask about the causes of disagreement over one problem at a time. Instead, I have suggested that we develop a taxonomy that lets us understand the type of
issues at stake. Then, once we know what kind of disagreement something is, we can better draw on our understanding of that type of disagreement to better develop and test hypotheses about how to resolve them.

I think that one way to more effectively adjudicate philosophical disputes (when it is possible to do so) is simply for interlocutors to ask each other the right questions to establish the rules of the game. My project aims to identify the fault lines in disagreement. In this chapter, I model judgments of moral responsibility and free will that were collected in an experiment, in order to identify a relationship between two philosophical concepts that, I argue, was not previously understood.

In this chapter, I use this approach to argue that free will is a nominal construct developed and deployed post hoc in an effort to provide cohesive narratives in support of a priori moral-judgmental dispositions. In a reversal of traditional course, I defend the view that there are no circumstances under which attributions of moral responsibility for an act can, should, or do depend on prior ascriptions of free will. Conversely, I claim that free will belief depends entirely on the apperceived possibility of moral responsibility. Orthodoxy dictates an agency-first thesis, according to which free will is necessarily antecedent to moral responsibility. However, I present a number of arguments against this view, and in favor of an agency-last stance, according to which the concept of free will is dependent upon that of moral responsibility.

I provide further support for my case in the form of new empirical evidence regarding the stable mode of inference used to attribute free will across moral contexts. These experimental results can be interpreted to imply the deflation of one of the longest-standing veridical paradoxes in experimental philosophy. Furthermore, the sole conceptual scheme found to be
capable of modeling the experimental results is also capable of illuminating several classic works in the analytic philosophy of moral agency.

4.1. A Puzzle of Moral-Agentive Cognition

On February 17th, 2004, the state of Texas administered a lethal injection to Cameron Todd Willingham for pouring gasoline all over his home, blocking the exit with a refrigerator, and setting fire to the house while his three daughters slept inside. Multiple investigators reported evidence of “mineral spirits” and “puddle patterns all over the place” from lighter fluid, which led from under the girls’ beds to the front door of the house. Willingham’s own lawyer has since said that over twenty pieces of “evidence showed that he was one hundred percent guilty” (Grann, 2009). One neighbor testified that as the fire blazed, Willingham calmly moved his car down his driveway to keep it away from the flames, and waited until the authorities arrived to “put on a show,” purportedly feigning a level of emotion that led a police chaplain to conclude that Willingham “was in complete control.” Supposing that these events were the inevitable result of natural laws, the invariable unwinding of a clockwork universe—that they took place in a universe that adheres to the philosophical thesis of causal determinism—might we have to relinquish the thought that Willingham really was in control of his behavior, that he even had the choice to behave otherwise?

The variance of philosophical beliefs is evident not only in interpersonal disagreements, but also in apparent contradictions of belief within a single subject. In theory, the sources of this within-subjects variance are manifold. In practice, one source of this apparent affinity to self-contradiction is order effects. Order effects are divergent responses to a single prompt,
depending on whether or not certain other philosophical questions were posed before it. Order effects seem to affect the beliefs of philosophy professors as much as, if not more than, the judgments of professors who are not philosophers (Schwitzgebel and Cushman, 2012). Moreover, it is not just that professional philosophers are as prone to these influences everyone else—philosophers who claim expertise about a particular subject are apparently just as likely to fall prey to these effects as anyone else is (Schwitzgebel, 2014). In the this chapter, I return to the idea that it is not individuals’ views we need to question as a result of observed covariance with individual difference factors, but what the underlying theories themselves have to tell us.

4.2. Motivating the Concept of Agency & the Concept of Motivated Agency

Why do philosophers care about free will? The impetus to preserve or willingness to forgo a viable notion of agency is generally motivated by a priori attitudes toward the need to maintain or ability to do without a robust concept of moral responsibility. But this only shifts our inquiry to an analogous one about why people care about moral responsibility. This shift suggests that moral responsibility may be a more fundamental concept than free will, which in turn raises important new research questions for metaphilosophy and moral psychology.

Normally, people assume that moral responsibility presupposes agency, but that picture may be perfectly backwards. Instead of facilitating meaningful blame and praise, free will may only be an instrument of blame and praise, a secondary and inessential construct through which we rationalize immediate dispositions to seek retribution and yield to obligation. This is roughly the opposite of what I have been taught, and I initially found the idea nearly impossible to fully wrap my head around. Nonetheless, I have found this perspective to yield a great deal of
explanatory power. If we consider that free will may only be a vestigial heuristic for explaining moral responsibility in the absence of more sophisticated neuropsychosocial explanations—rather than a necessary condition for formally establishing moral responsibility—then it becomes unclear if the existence of free will is relevant to moral inquiry at all.

4.2.1. Topical and Methodological Purview

In §4.3, I discuss a widely studied puzzle of moral-agentive cognition,\(^{10}\) which I use as a window into the current moral psychological landscape. Against this backdrop, I contrast my own views on moral-agentive judgment in practice (psychological questions) and in theory (philosophical questions). I identify in §4.4. a number of matters arising from previous attempts to understand the puzzle, and offer a way to circumvent the logical inconsistencies encountered on the approaches traditionally taken by philosophers and psychologists. I do so by proposing a change in perspective, which I more fully outline and compare to its theoretic competitors in §4.5.. In §4.6., I take a decidedly empirical turn, in order to test the predictive accuracy and relative fit of each theoretical paradigm\(^{11}\) of explanation. In that section, I operationalize the proposed new

---

\(^{10}\) By moral-agentive cognition, I mean the processes through which the moral agency of others is evaluated. I deliberately use the opaque term ‘moral agency’ and its conjugates in all propositions that refer ambiguously to moral responsibility and free will, but which fail to distinguish between the two. My reasoning for employing this uncommon catchall is to flag and avoid the potential pitfalls addressed in §4.5.

\(^{11}\) I use the term ‘paradigm’ in the fourth sense listed in the *Oxford English Dictionary*: “A conceptual or methodological model underlying the theories and practices of a science or
Factlessness & Faultlessness  

Holtzman 61

approach, as well as its theoretic competitors, as predictive models of moral-agentive cognition, and test each of these models against experimental data of ordinary philosophical judgments. The discussion of these results, which constitutes §4.7., serves to explain of how the model I defend can reduce to triviality several psychological pseudoproblems arising from the purportedly “paradoxical” (Sinnott-Armstrong, 2008) nature of ordinary beliefs about moral responsibility and free will.

Finally, I discuss in §4.8. how the prima facie renegade theory I defend is in fact anything but the radical departure from tradition that it may at first appear to be. On the contrary, the agency-last paradigm I embrace is uniquely capable of providing a unified philosophical perspective from which we can more easily develop a coherent picture of several of the most prominent metaethical arguments in 20th century philosophy. Rather than debating whether or not free will exists, my interest here is in considering the question of whether, from a strictly moral perspective, free will matters. My thesis is that regardless of whether or not free will exists, its existence—something about which we can neither have certain knowledge nor certain doubt, and which we could never come to know through experience (Kant, 1785/1998)—may be largely irrelevant to the domain of moral inquiry, the domain with which most people who entertain questions of agency are primarily concerned.

discipline” (2014). I reserve the word ‘model’ for discussions of statistical tests of observed moral-agentive attributions, which I analyze with a technique called structural equation modeling (Kline, 2011).
4.3. A Puzzle of Moral-Agentive Cognition

Why do we tend to view others as autonomous agents—as endowed with the freedom of choice (Nahmias 2006), as behaving intentionally (Knobe, 2004), and as possessed of causal powers (Alicke, Rose, and Bloom, 2011)—in proportion to the extent to which they have ‘done wrong’? With greater moral transgressions comes greater—or at least more certain—moral responsibility. This stands to reason, as the responsibility one shoulders by killing thousands in cold blood far exceeds the moral burden undertaken by lighting up a joint or smoking a cigarette in a no-smoking zone. But does it also stand to reason that greater moral transgressions are indicative of greater levels of volition, deliberation, and instrumentality?

4.3.1. Type and Token Thought Experiments

Consider a deterministic universe, in which...

...scientists figure out the exact state the universe was in at the time of the big bang, and figure out all the laws of physics as well. They put this information into a supercomputer, and the computer perfectly predicts everything that has ever happened and ever will happen. In other words, these scientists prove that everything that happens has to happen exactly that way because of the laws of physics and everything that's come before.

Now, suppose that in such a universe…
...someone commits a crime.

With both these premises in mind, consider two questions:

Was this person free to choose\textsuperscript{12} whether or not to commit this crime?

How morally responsible is this person for committing this crime?

Next, consider a second scenario, adapted from the actual case of Cameron Todd Willingham, the triple-murder-arson suspect who ultimately received the death penalty for the crimes of which he was accused (Grann, 2009). Suppose that in this same deterministic universe...

...a man named Todd has taken to abusing his daughters. In order to cover up this abuse, he pours gasoline all over his home one morning while his wife is out shopping, lights the house on fire, and successfully murders his daughters while remaining unharmed himself.

Here, we can ask:

\textsuperscript{12} Throughout this chapter, I treat ‘free will’ and ‘free choice’ as interchangeable. If this gives the reader the impression that I (or the participants in the experiment) do not know how one is supposed to use the term ‘free will,’ then the reader is beginning to get my point.
Was Todd free to choose whether or not to commit this crime?

How morally responsible is Todd for committing this crime?

4.3.2. The Source of the Puzzlement

Even when asked to assume determinism, people are more likely to make attributions of moral responsibility (‘MR’) (Nichols and Knobe, 2007) and of free will (‘FW’) (Nahmias, Morris, Nadelhoffer and Turner, 2005) to agents who have committed more severe moral transgressions. Especially puzzling is the fact that for some types of act (e.g., crimes), most people tend to deny MR and FW, yet they usually assert MR and FW for certain token acts of those very same types (e.g., Todd’s criminal infanticide). Previous researchers have generally defended one of two interpretations of the effects of the affective, concrete, or morally transgressive nature of an act (‘ACT’) on attributions of moral agency.

4.4. Normative and Deflationist Accounts of the Puzzle

What do people really believe about the compatibility of free will and causal determinism, and

---

13 I use MR, FW, and ACT to refer to the psychological constructs, rooted in personal perception, of which participant ratings are an indicator. In light of recent evidence and arguments for ‘interpretive diversity’ (Nichols and Ulatowski, 2007), I think it is important to distinguish these constructs from the potentially mind-independent, objective, real phenomena with which philosophers are typically concerned.
are their beliefs contradictory? Historically, experimental philosophers have adopted what I will call the *normative approach*, taking results like these to show that concrete, affect-laden details can bias attributions of moral agency (as mentioned in §2.2.1.). Because these normative theorists believe that these results reveal a logical contradiction in ordinary attributions of moral agency across contexts, many researchers have taken such studies as evidence of a kind of “abstract/concrete paradox” (Sinnott-Armstrong, 2008). From this point of departure, these researchers have sought to develop *error theories*. Error theories are characterizations of the *situational factors* that select for faulty processing and attribution of moral agency (Leslie, Knobe, and Cohen, 2006), and of the *cognitive processes* that lead to these errant judgments. But there is a rift between error theorists who think that *ACT* inhibits the production of competent, accurate attributions of moral agency, and those who believe that it actually facilitates accurate moral-agentive attribution.

One camp of normative theorists, perhaps guided by their own *a priori* commitment to the philosophical thesis that genuine moral agency is incompatible with determinism, insists that the intrusive presence of certain details leads people to erroneously attribute inflated levels of moral responsibility and free will. These expressed opinions, they argue, actually belie people’s true, “naturally incompatibilist” beliefs (Kane, 1999). Members of the other camp, who are sometimes motivated by an overt interest in promoting the belief that moral agency is compatible with moral responsibility and free will (Vohs and Schooler, 2008), argue instead that it is the

---

14 One exception to this rule is the *Norm Broken, Agent Responsible (NBAR)* theory (Ripley and Mandelbaum, 2012), which focuses on transgressions rather than emotion/cognition or abstract/concrete distinctions. Most of my criticisms do not apply to these sorts of value-neutral approaches, and I think that projects like these may be particularly informative.
absence of affective details that biases people, causing them to mistakenly mitigate the levels of moral responsibility and free will they assign in the presence of deterministic factors.

Opposed to the normative approach is the deflationist approach noted in §2.2.1,\textsuperscript{15} which is typically defended by skeptics about experimental philosophy. Such skeptics believe that experimental philosophy has nothing to contribute to philosophy as traditionally construed, and that empirical approaches can only reveal the unstable “folk intuitions” (Kauppinen, 2007) of philosophical novices who have been confronted with stripped down cases and denied sufficient time or resources to adequately reflect on them. While normative theorists find results like those discussed to be interesting and surprising, deflationists raise doubts about the philosophical import of such findings (Cappelen, 2012), and about the broader, potentially insidious assumption that “intuitions are likely to be reliable and should form the building blocks for sound moral judgments” (Sunstein, 2005). From a skeptical, if somewhat reductive perspective, such findings might be seen as nothing more than specific examples of the general principle that upsetting events influence ordinary judgments. If concrete, affective details only influence ordinary attributions of moral agency through domain-general processes—that is, through processes capable of biasing all sorts of cognition, not just moral-agentive cognition—then the effects of \textit{ACT} on \textit{MR} and \textit{FW} might not be of any uniquely moral or philosophical significance.

\textsuperscript{15} It would be misleading to contrast deflationists directly with error theorists. Many (and I suspect most) deflationists believe that philosophical questions have objectively correct and incorrect answers. On this view, competent and errant responses might be thought to correspond directly to these correct and incorrect answers, respectively. The primary difference between (realist) deflationists and normative theorists, then, is only the philosophical import each attaches to her preferred error theory.
4.4.1. Shortcomings of the Normative and Deflationist Approaches

But neither the normative nor the deflationist approach is entirely well-founded, and both interpretations impugn the so-called “folk intuitions” of outsiders to the analytic philosophy community without sufficient justification for doing so. Each approach purports to tell us what to make of the variance in harshness and leniency with which people make moral-agentive attributions in different situations. But it does not follow from the fact that in different situations people tend to respond differently (more or less harshly or leniently) in their moral-agentive attributions, that in some of these situations they tend to respond differently than they should (too harshly or leniently). To claim that instability in the application of agentive concepts across cases can properly be characterized as logical inconsistency in those concepts’ application is to make two unwarranted assumptions.

First, there seems to be no way of deciding which of these two ostensibly inconsistent applications is errant, other than by reference to one’s own a priori attitudes toward the compatibility or incompatibility of moral agency with causal determinism; the ‘error theory’ to which a researcher subscribes will inevitably be biased by her own philosophical beliefs. The reason for this is straightforward. The assertion that a person has responded differently than she should have implies certain propositions about the range of responses she could have given which would have been more acceptable. But these propositions amount to little more than claims about whether or not full causal determinism may be the sole factor that decides moral agency, even in the presence of other moral and agentive factors. As such, any theory premised on an assertion of inconsistency (as opposed to mere instability) would have to presuppose the very same highly contentious philosophical framework it sought to crown as the one most people
really believe—either compatibilism, or incompatibilism.

As an illustrative example, imagine that an empirical scientist wishes only to defend a ‘psychological reading’ of one of these error theories, which only describes behavior and the component processes leading to it. This scientist, we can assume, wishes to eschew anything that might be construed as a ‘philosophical reading,’ a reading that extends interpretation to the metaphysical or normative-ethical domains. In other words, she intends only to describe the processes by which people arrive at different philosophical conclusions in different cases, and to indicate the goal-directed success or unsuccess of those processes in each case. Still, her error theory must characterize certain practical judgments as either representative or non-representative of subjects’ competent, unbiased beliefs, and therefore as resulting from cognitive processes taking place under ideal (controlled) or non-ideal (confounding) conditions. But characterizations of this latter kind require her to embrace some set of a priori views toward the relevance or irrelevance of certain ACT details to competent, unbiased moral-agentive judgment. Such views are tantamount to a priori assumptions about the normativity or deviancy of the circumstances under which these acts are committed, and in which these moral-agentive assessments are made. From this, it follows that some of our researcher’s a priori assumptions will have to be ones about the moral normativity or deviancy of the acts committed and judged.

Second, in taking the instability of moral-agentive ascriptions across situations to be logically inconsistent, researchers usually attribute these supposed mistakes to the presence or absence of emotional or affective processes. But the tacit assumption that the key factor is affective or other processes per se, rather than the component of affect (or other aspects of mentation) central to and perhaps even constitutive of (Prinz, 2007) judgments of moral agency, is unwarranted. If it turns out that the details of a moral transgression (or the way in which that
transgression is presented) affects either MR or FW in a way that is essentially indirect—that is, which only occurs due to some causal link between MR and FW—then studying these indirect effects could reveal fundamental facts about the conditional relationship between the ordinary language concepts of moral responsibility and free will. Such findings could reveal attributions of free will to be largely inert in the initial assignment of moral culpability. If this were true, we would have reason to doubt that agentive ascriptions are anything more than moral honorifics rendered ex post facto.

4.4.2 The Need for A New Approach

Ultimately, the philosophical project of deciding which moral-agentive evaluations are appropriate in a given situation is inseparable from the project of deciding whether a given situation is conducive to making appropriate moral-agentive evaluations. In order to develop any error theory, a researcher must make a priori assumptions about the liberty or illiberty of agents, and about the moral normativity or deviancy of acts committed and judged. The unjustified and, I have argued, seemingly false presuppositions made by both kinds of error theories necessitates an alternative explanation of the effects of ACT on moral-agentive cognition. This demands that we shift our research focus away from trying to decide which judgments of moral agency are right and which are wrong, and turn instead to a new set of questions about the relationship between agency and moral responsibility.

The relationship between moral responsibility and free will is usually taken to be analytic (that is, true by definition), so focusing on the nature of the logical connection between these concepts is not only useful from a philosophical perspective, but might also help us develop and
test a psychological model of judgments of moral responsibility and free will. This, in turn, might tell us something about the analytic interdependence of these two philosophical concepts. Rather than inquire as to the individual differences that underwrite moral and philosophical disagreement and misunderstanding as in Chapter Two (Holtzman, 2013), my purpose here is to explore an even more perplexing question. This is the question of whether we first assess free will, and on the basis of this information build our understanding of an agent’s moral responsibility, or if we first make attributions of moral responsibility, and only later construct agentive narratives on the basis of these attributions.

4.5. Paradigms of Moral-Agentive Relata

What are the common conceptual schemata that allow us to coherently debate and discuss such opaque phenomena as liberty, willfulness, and agency? Here, I begin to address the focal problem of how it is possible to infer moral responsibility and free will from ordinary observations of behavior. In particular, I am interested in what causes (or allows) people to draw these inferences even when they assume causal determinism (one version of which I have laid out in §4.3.).

The thesis that causal determinism is compatible with free will is, for reasons that may be obvious, referred to by philosophers and psychologists as compatibilism; and the thesis that the two are not compatible is called incompatibilism. For reasons that may be equally obvious, the thesis that causal determinism is compatible with moral responsibility is also referred to as compatibilism, and the thesis that the two are not compatible is called incompatibilism. But as illustrated by way of parable in Frankfurt (1969), and explicitly elucidated by Fischer (1982),
compatibilism and incompatibilism about causal determinism and moral responsibility are distinct from compatibilism and incompatibilism about causal determinism and free will.

In coining the terms *compatibilism* and *incompatibilism* (which arose in §2 and in §3.2.4.), Slote (1969) intended them to refer to only the theses that we might call *agentive compatibilism* and *agentive incompatibilism*. For Slote, these were views about free will, which did depend on “one’s evaluation of certain moral issues,” but also on a number of other factors, including “the force and significance of certain similes, analogies, and diagrams.” They are broader versions of the compatibilist theory that William James (1884) called *soft determinism*, and the incompatibilist position he dubbed *hard determinism*, more inclusive because they do not require the assumption that determinism is true (an assumption which, it just so happens, James rejected). But in time, Slote’s words have replaced Strawson’s (1963) *optimism* and *pessimism* and begun to lead a double-life, moonlighting as shorthand for *moral compatibilism* and *moral incompatibilism* while keeping their jobs as monikers for *agentive compatibilism* and *agentive incompatibilism*. It should be clear, then, that any thoroughgoing understanding of “Folk Intuitions About Moral Responsibility and Free Will” (Nahmias *et al.*, 2005) must recognize the distinction and relationship between these two kinds of compatibilist thesis—ones concerning free will, and ones concerning moral responsibility.\(^{16}\) This recognition naturally requires us to understand the distinction and relationship between free will and moral responsibility. The

\(^{16}\) It should be acknowledged that Nahmias *et al.*'s 2005 paper, contemporary reactions to which kicked off the experimentalist debates between compatibilists and incompatibilists, did in fact measure both constructs. It may be due in part to the unfortunate similarity in ratings of *MR* and *FW* for the particular cases tested by Nahmias and his colleagues that researchers have since begun to assume that the two types of attribution are more-or-less identical.
4.5.1. The Identity Paradigm

In their watershed paper, “Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions,” Nichols and Knobe warn that “one might maintain that determinism is compatible with moral responsibility but not with free will” (2007). But with rare exception (e.g., Feltz, 2013), it has become status quo for psychologists and philosophers to treat ‘free will’ as synonymous with ‘moral responsibility’ (e.g., Baumeister, 2008; Paulhus and Carey, 2011 Rose and Nichols, 2013; Feltz and Cokely, 2008;). While this identity paradigm may be expedient for communicating research to the public in the sexiest way possible, it makes a mess out of some of the best-known metaphysical arguments of the last fifty years. The issue is not merely that the identity paradigm stands in contrast to certain philosophical positions that some of us might like to defend; the issue is that the paradigm itself renders entire debates, and every contrasting position within those debates, utterly incoherent.

Fischer’s semicompatibilism, developed from the premise that we can “separate compatibilism about causal determinism and moral responsibility from compatibilism about causal determinism and freedom to do otherwise” (1987), would be far less influential if he had endeavored on the identity-paradigmatic project of trying to separate compatibilism about moral responsibility from, well, compatibilism about moral responsibility. Van Inwagen’s assertion that we possess “the free will required for moral responsibility” would be a fallacious case of affirming the consequent, were we to think of him as arguing that people have the moral responsibility required for moral responsibility (1983). And Pereboom’s denial of “whatever sort
of freedom is sufficient for moral responsibility” (2001) would read as miserably circular to believers in free will, if all he really was denying was the sort of free will sufficient for free will.

Philosophers have done an excellent job of identifying the ways in which moral responsibility and free will are theoretically orthogonal. As experimentalists work to identify the ways in which attributions of these concepts are made, it is important not to lose sight of their theoretical distinction, as this distinction could be expected to dictate their use in practice. Any satisfactory account of moral agency or the attribution thereof should recognize this distinction. Therefore, one of three possible non-identity accounts can be expected to provide greater insight than the identity paradigm.

4.5.2. The Common Causes Paradigm

One possibility is that attributions of free will and moral responsibility are independent and immediate responses to certain kinds of perceived acts. On this view, their covariance (or tendency to co-occur) belies their common dependence on some third variable or set of variables. The puzzle at hand suggests that the culprit, should this view be correct, is their shared ancestry in the affective, concrete, or morally transgressive nature of an act. Within this common causes paradigm (Figure 10a), MR and FW are not conditional on one another. Unfortunately, this approach raises a glaring question: If moral responsibility and free will are not causally linked, why should affective, concrete, and moral factors influence judgments of free will?

The common causes paradigm demands an independent path for this influence, but it is not clear why we should expect such a path to exist. It seems reasonable that a person’s moral responsibility at some time $T_2$ depends on just what she has done at some earlier time $T_1$, but it
does not seem possible that the free will an agent possesses at some earlier time $T_0$ could depend on what she has not done and will not do until some later time $T_1$. The common causes paradigm, it can be seen, unyokes moral responsibility and free will too much.

![Figure 10](image)

**Figure 10.** Three competing paradigms of moral-agentive entailment relations. (A) The *common causes paradigm* takes the correlation between $MR$ and $FW$ to be a spurious “third variable” phenomenon. (B) The agency-first paradigm assumes that free will is assessed prior to judgment of moral responsibility, and therefore can affect its assessment but not vice-versa. (C) The agency-last paradigm, advocated here, is based on the theory that judgments of moral responsibility are immediately influenced by the details of moral transgressions, and attributions of free will are essentially conditional on these earlier judgments of moral responsibility.

4.5.3. The Agency-First Paradigm

Most philosophers have endorsed, at least implicitly, an agency-first paradigm (Figure 10b). This approach introduces a level of parsimony by treating $MR$ and $FW$ as serial rather than parallel judgments, which may be appropriate insofar as moral responsibility and free will are serial phenomena. By definition, an agent’s free will at some time $T_0$ can only play a causal role in an event which occurs at some later time $T_1$, and she cannot be morally responsible for the
consequences of that event until some even later time $T_2$, once those consequences have manifest. In keeping with this sequence of events, the *agency-first paradigm* contends that the association between attributions of moral responsibility and of free will is due to a conditional relationship thought to hold between them: To whatever extent $ACT$ informs $FW$, a portion of these effects will be indirectly transmitted to $MR$ via their influence on $FW$. Because this sort of conditional process in moral-agentive cognition is perceived by many to be a desirable reflection of the relationship between the metaphysical concepts of moral responsibility and free will, it is the relationship most commonly assumed by social scientists (Baumeister, 2008) and philosophers (Dennett, 1984) to hold between $MR$ and $FW$. But this assumption is based on fallacious reasoning.

Belief in the priority of agency over responsibility is driven primarily by commitment to a widely-held dogma regarding temporal priority, not logical antecedence. This dogma dictates that an agent must freely choose to commit an act prior to its occurrence, and therefore prior to being responsible for its occurrence, in order to be morally responsible for that act. While this may be true, logical priority and temporal priority are two very different matters, and in the case of necessary preconditions, the arrow of time and the arrow of entailment will always head in opposite directions. To say that moral responsibility must follow free will is to say that free will must follow from moral responsibility. Therefore, the fact that a person must freely choose to commit an act before she can be morally responsible for its consequences does not license us to infer moral responsibility from free choice; instead, the fact that a person is morally responsible licenses us to infer that she has acted freely. Philosophers who wish to maintain that free will is a necessary precondition for moral responsibility are therefore in no position to defend the *agency-first paradigm*. 
On the contrary, philosophers have unwittingly developed a conception of moral responsibility and free will that implies an agency-last paradigm (Figure 10c). The potential philosophical implications of the suggestion I have just made are not insignificant. The concept of moral responsibility plays, at the very least, an important causal role in individual and group behavior. For this reason, even those who believe it to be little more than an instrumental construct have reason to pay it heed, in the interests of social regulation and social explanation. Free will is also thought to play a role in social cognition, in substantiating attributions of moral responsibility. But the agency-last paradigm suggests that whatever roles free will might play in social cognition, establishing moral responsibility from the get-go is not one of them. If so, its deployment ex post facto as a source of moral justification can only underwrite circular arguments.

4.5.5. From Psychological Models to Conceptual Paradigms

Traditionalist readers, who might be expected to embrace the agency-first paradigm and perhaps also to eschew metaethical naturalism (see Prinz, in preparation), may at this point be unimpressed by the claims I have just made about inference. In fact, some readers sympathetic to my views will have recognized that inferential priority is not the same thing as logical priority. But because people generally take the relationship between free will and moral responsibility to be analytic—and in particular, because philosophers who care about free will will justify their interest
on the basis of their analytic relationship—inferences about moral agency are strictly logical ones.

In the substantive sense in which many philosophers aspire to speak about free will, a person cannot, *ex post facto*, choose past courses of actions for which she is already morally responsible, anymore than she might foresee past events for which others were morally responsible. Conversely, a person can only be known by others to have partaken in some event of her own free will *ex post facto*, just as foresight can only be confidently attributed in retrospect. Thus, the assumption that free will can be established independently from moral responsibility is based on a gross misunderstanding of the basic logic of moral causation. Philosophers who wish to maintain that free will is a necessary precondition for moral responsibility are therefore the last people who should want to defend the *agency-first paradigm*.

If free will has no empirical correlates in daily life, no explanatory role in theory, and no instrumental value in social cognition, then it is unclear why we should concern ourselves with that concept at all. The purpose of the next section is to test the hypothesis that attributions of free will are secondary to judgments of moral responsibility, and do not play an essential role in the understanding or regulation of normative judgment.

### 4.6. Experiment

From where do we derive the apperception of free choice? Here, my goal is to show that the variance in *FW* across moral transgressions can be accounted for entirely by the variance in *MR* across those transgressions. I also hypothesize that, conversely, the variance in *FW* fails to explain a significant amount of the variance in *MR* across cases. Together, these two findings
would show that $\text{MR}$ and $\text{FW}$ are distinct, yet not entirely independent judgments. Moreover, such findings would demonstrate that the influence of $\text{ACT}$ on $\text{FW}$ can be understood entirely in terms of moral considerations of a transgression, so long as we accept the view that free will falls out of—rather than factors into—moral responsibility. To test the predictive accuracy and relative fit of each paradigm, undergraduates were recruited to participate in an experiment.

4.6.1. Participants

Participants were selected from introductory philosophy classes in the university-wide required core curriculum at Brooklyn College, and were told that participation would not affect their course grades. Responses were delivered with no identifying information, and pen-and-paper surveys were proctored by a professor who was not their instructor, in order to preclude any form of actual or perceived coercion. All instructors of classes from which participants were recruited indicated that free will, moral responsibility, determinism, and compatibilism had not yet been addressed in their courses. All students who passed a comprehension check$^{17}$ and filled out every response of interest were included in the analysis ($n = 228$). Of these, 42% were female; 5.3% over age 30; 46% White, 19% Asian, 17% Hispanic or Latino, 10% Black, and 8% Other or =

---

$^{17}$ After reading about a deterministic world, participants were asked whether our world is subject to the very same deterministic principles. Participants who responded ‘Yes’ to this question, and who also indicated that the agent in the vignette was ‘Not at all responsible,’ were excluded from all analyses. Presumably, these participants (a) did not understand the task, (b) did not take the task seriously, or (c) rejected the very concept of moral responsibility $a$ priori, regardless of the (as-yet unread) case at hand.
First, participants were asked to imagine the deterministic universe described in §4.3. To check that between-group differences in MR and FW were due to the experimental manipulation rather than a priori differences in willingness or ability to assume a deterministic universe, participants were asked whether they thought that our own world is deterministic in a way similar to the world they read about. Overall, 17.5% of participants indicated that our own world is deterministic in the same way as the world described in the vignette. There was no significant difference in rate of a priori determinism between groups ($t_{226} = .347, p = .729$).

Next, participants were presented with one of the two crime vignettes described in §4.3. Participants were then asked to indicate whether or not the agent freely choose to commit his crime, and were then prompted to rate the agent’s level of moral responsibility on a Likert-type scale from 1 (“Not at all responsible”) to 7 (“Absolutely responsible”). These procedures again used the materials presented in §4.3. FW was always collected first, in order to increase the likelihood that participants would evaluate free will before considering moral responsibility, thereby increasing the probability that the proposed agency-last model would be rejected.

4.6.3. Results

Replication. Initial tests successfully replicated previous findings in the literature. As in these earlier studies, there was a significant relationship between the vignettes participants read, and the levels of moral responsibility they assigned ($t_{226} = 9.64, p < .001$). Participants who read

---

18 Vignette materials are presented verbatim, in offset, italicized text in §4.3.1.
about the gruesome triple-murder-arson tended to view its perpetrator as much more morally responsible for his actions ($M_{114} = 6.535$, $SE = .116$) than participants who were asked to judge a generic criminal ($M_{114} = 4.404$, $SE = .188$). Participants who read of this more heinous crime also demonstrated significantly higher odds (OR = 2.74) of accusing its perpetrator of freely choosing to commit his crime ($\chi^2_1 = 10.058$, $p < .005$).

**Differentiation.** Further analyses supported the hypothesis that moral responsibility and free will were treated as distinct concepts (Figure 11). When participants were collapsed across experimental conditions, there was a medium-sized (Cohen, 1988) correlation between $MR$ and $FW$ ($r = .445$, $p < .001$). But as anticipated, the relationship between these ascriptions was moderated by the vignette participants read, $F_{1, 224} = 6.313$, $p < .05$. Although their correlation was medium-sized among participants who read about the child-killer ($r = .486$, $p < .001$), it was small for those who read about a generic criminal ($r = .247$, $p < .01$).

**Mediation.** A path analysis (Hayes, 2013) was then conducted to test the hypothesis that $MR$ mediated the effect of $ACT$ on $FW$.\(^{19}\) This prediction was supported by the finding that there remained no significant main effect of $ACT$ on $FW$ (Figure 2). To rule out possible mediation in the other direction, a second path analysis was conducted against the more traditional view that $FW$ mediated the effect of $ACT$ on $MR$. This analysis found no evidence of *agency-first* mediation. As predicted, the direct effect of $ACT$ on $MR$ remained significant ($z = 9.01$, $p < .001$) even after accounting for potential indirect effects of $ACT$ on $MR$ (Baron and Kenny, 1986).

---

\(^{19}\) Coefficients of dichotomous and continuous variables were made comparable by the method first recommended by MacKinnon and Dwyer (1993), and expanded upon by Hayes (2009).
Figure 11. Parameter estimation. Differences in free will attribution between conditions were totally mediated by judgments of moral responsibility. After controlling for differences in MR, participants who read details of a grotesque arson actually made slightly less frequent attributions of free will than those who only read only of a generic crime, though this effect did not approach statistical significance ($p = .463$).

Evaluation. Finally, fully-specified agency-first and agency-last models were tested, and were compared to one another and to a common causes model (Table 5). Only the agency-last model predicted a pattern of responses that did not differ significantly from the actual data collected (Likelihood-ratio $\chi^2$). That model was also the only one to provide an acceptably close absolute fit to the data after adjustments for parsimony were made to each model (RMSEA; Brown and Cudeck, 1993). Comparisons between models were then made with a parsimony-adjusted index of relative fit, the Bayesian Information Criterion (BIC). These comparisons found that the agency-last model provided a better relative fit than either of the other two nested models, and was an even better fit than a just-identified model (Kline, 2011) in which all variables were correlated with one another (BIC = 1459.405). The agency-last model was also the only one to
provide an acceptable improvement in fit over a null model (CFI; Hu and Bentler, 1999). The common causes and agency-first models both left unacceptably large correlations between variables unexplained (SRMR), but on the agency-last model, less than 1% of the unaccounted for (residual covariance) between ACT and FW could have been due to a common factor.

<table>
<thead>
<tr>
<th></th>
<th>Agency-First</th>
<th>Common Causes</th>
<th>Agency-Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>69.403*</td>
<td>41.25*</td>
<td>.538 (n.s.)</td>
</tr>
<tr>
<td>RMSEA</td>
<td>.548*</td>
<td>.420*</td>
<td>.000 (n.s.)</td>
</tr>
<tr>
<td>BIC</td>
<td>1523.379</td>
<td>1495.226</td>
<td>1454.513</td>
</tr>
<tr>
<td>CFI</td>
<td>.459</td>
<td>.682</td>
<td>1.0</td>
</tr>
<tr>
<td>SRMR</td>
<td>.150</td>
<td>.112</td>
<td>.012</td>
</tr>
<tr>
<td>Residual covariance</td>
<td>.443</td>
<td>.285</td>
<td>-.008</td>
</tr>
</tbody>
</table>

Note: * $p < .001$

Table 5. All fit statistics favored the agency-last model, and disfavored the common causes and agency-first models. $\chi^2$, RMSEA, and BIC are measure badness-of-fit, for which higher values indicate poorer fit (more unexplained covariance). For these measures, $\alpha$ levels indicate the probability that the model will correctly predict the responses of a randomly selected population sample. The comparative fit index (CFI) measures goodness-of-fit, with higher values indicating better fit between the model and the data. This measure reflects the marginal improvement in prediction each model provides over a model assuming a random distribution of responses. The
SRMR provides a measure of the mean absolute value of the unexplained covariance between each possible pair of variables, and the residual covariance measures the association between those two variables for which no path is specified on the model.

4.7. Discussion

How do people actually think about moral responsibility and free will? The experiment suggests two clear answers to this question. First, people generally conceive of moral responsibility and free will as related but decidedly different constructs. In the sample-population, attributions of free will were much less susceptible to the influence of moral and emotional considerations than were judgments of moral responsibility. Put another way, the extent to which MR and FW came apart depended on differences in the acts under consideration. Because these two moral-agentive concepts diverged in this way, it can be inferred that they are not identical. Therefore, any account of how people employ either psychological construct cannot be successful without sufficiently distinguishing the two.

Second, judgments of free will across moral contexts appear to be driven primarily by evaluations of moral responsibility. No aspect of the acts with which participants were confronted, other than those acts’ tendencies to elicit attributions of greater or lesser moral responsibility, significantly affected beliefs about the freedom of choice exercised by their perpetrator. Importantly, this means that attributions of free will and moral responsibility came apart in precisely the way that an optimist about the ordinary ability to recognize and to distinguish between moral responsibility and free will might hope they would.
When modeled according to these two principles, ordinary attributions of free will did not appear to be significantly affected by any form of concrete or affective bias. This discovery stands in stark contrast to previous characterizations of these attributions as noisy and irrational. On the basis of these findings, I will now sketch a series of responses to the questions I set forth earlier. I expect that my account will be incomplete and not entirely precise. The discussion that follows is intended primarily as a framework within which to develop future research programs bridging conceptual analysis with the study of social cognition.

4.7.1. From where do we derive the apperception of free choice?

There appear to be at least two ways in which people become inclined to view others as possessed of the freedom of choice. The first route to attributions of agency may at first seem ineluctable: unless there is strong evidence to the contrary (and often despite strong evidence to the contrary), the free will of others may often be taken for granted. I will come back to this route in a moment. The second positive route is through the ascription or perception of moral responsibility. Other well-trod explanatory roads to agentive attribution, including ACT, may only be indirect tributaries to ascriptions of agency, contributing only as an indirect result of their more immediate connection to moral responsibility.

To be clear, those who have wondered why ACT bears on FW have not been mistaken about the existence of such a relationship per se, but have failed to recognize the chain of influence through which this indirect effect is transmitted. A framework that recognizes this chain of influence is uniquely capable of supporting a psychologically and philosophically satisfying explanation of these conditional effects. ACT does not influence MR and FW so much
as it influences \textit{FW because and only because} it influences \textit{MR}. This chain of influence may be difficult to see because it contradicts the deeply-entrenched \textit{agency-first paradigm}. The \textit{agency-last paradigm} remedies this error, and the reinterpretation of the data it makes possible should be seen as a victory by all those who believe that genuine philosophical concepts and their logical employ are accessible to most people. But how might we account for the tendency to default to the \textit{agentive assumption}, even in non-moral contexts?

The \textit{agency-last paradigm}, when more fully developed, may actually allow us to understand and study the seemingly ineluctable route to free will beliefs as well. For example, one might wonder about the extent to which freedom of choice plays a central role in cultural identity, and to which cultural identity informs beliefs about freedom of choice. In recent work, Prinz (in preparation) has noted that the tendency to automatically associate personhood with agency is very likely learned. If so, and if the \textit{agency-last} hypothesis is correct, then it might be expected that socialized differences in moral responsibility norms should produce corresponding differences in free will belief. On the other hand, if moral responsibility presupposes free will, then fatalists should presumably partake of less retributivist forms of punishment, since only non-moral, practical concerns should guide their punitive decisions. To test this hypothesis, Prinz looked at the penal systems of the ten most fatalistic countries in the world (according to the World Values Survey; Minkov, 2012). Although less than one-third of all countries have retained capital punishment in law and practice, eight of the ten most fatalistic countries in the world today are retentionist states. While these findings border on anecdotal for the moment, they are suggestive of the kind of research impact that might be achieved in philosophy and psychology by flipping Kant on his head in the way suggested by the \textit{agency-last paradigm}. 
4.7.2. What are the common conceptual schemata that allow us to coherently debate and discuss such opaque phenomena as liberty, willfulness, and agency?

It follows from the non-identity of the psychological constructs of \( MR \) and \( FW \) that the philosophical concepts of moral responsibility and free will, which those constructs are meant to serve, are also distinct. This must be true whether or not error theories are appropriate for characterizing the variance of \( MR \) and \( FW \) across moral contexts. If we are to sacrifice error theories, as I have suggested we should, then we are plainly are dealing with two different concepts. But even if the effects of \( ACT \) on \( MR \) and \( FW \) might properly be accounted for by some error theory, then the full set of processing errors each is prone to would have to be different. It follows from this that the core competencies underwriting each must also be different. Because core competencies by definition involve the ‘right’ use of concepts, deflationists and normative theorists alike are forced to acknowledge that philosophical concepts of free will and moral responsibility, and the psychological constructs that underwrite their attribution are, at their core, distinct.

An objector might reply that the principles of deduction permit the inference of causes from their effects—or more precisely, the inference of necessary preconditions from the observation of phenomena that require the satisfaction of those conditions. One might therefore be tempted to think that the inference of \( FW \) from \( MR \) tells us nothing about the metaphysical relationship between moral responsibility and free will, or even about the priority of \( MR \) over \( FW \) in individuals’ conceptual schema. But the inference of a cause from its ostensible effects requires external evidence that the proposed cause exists. Without such evidence, theoretic
causes only play a nominal role in arguments that ultimately beg the questions they set out to answer.

For example, the existence of our atmosphere can be validly deduced from the continued observation of lightning on Earth; but so can the existence of Thor, if we assume *a priori* that Thor exists. And not only the equations of Special Relativity, but also the supposed existence of a luminiferous ether, was “deduced from the phenomena of light” (Maxwell, 1878). But the latter, unlike the former, was unjustifiably premised on the apparently false (and inherently etheric) assumption that light always “must be somewhere, and supported by some material agency” (Poincaré, 1908). Likewise, the inference of free will from moral responsibility requires the further, unwarranted assumption that moral responsibility must be mediated by some metaphysical agency—that is, that that free will exists in the first place.

Against the presupposition that there exists a material ether that mediates the transmission of light between distant bodies, John Stuart Mill once pointed out that “cases may be cited, even in our imperfect acquaintance with nature, where agencies that we have good reason to consider as radically distinct, produce their effects, or some of their effects, according to laws which are identical” (1868). This same point can be applied to the presupposition that free will exists. Just as the revolution in physics at the turn of the 20th century eventually enabled a shift away from ethereal theories of luminescence, recent advances in neuropsychology, genetics, and other human sciences at the turn of the 21st century have begun to obviate the postulation of free will.

4.7.3. *What do people really believe about the compatibility of free will and causal determinism, and are their beliefs contradictory?*
There is nothing it could mean to discover that most people are “natural” (as opposed to “unnatural”) compatibilists or incompatibilists. This is not just due to the fact that most people have never thought about compatibilism, or that they (like Strawson) “do not know what the thesis of determinism is” (1963), or that they (like I) do not know what free will is supposed to be. It due to the fact that any claim about whether or not most people are “natural compatibilists” will rest on an error theory, and this will in turn rest on theoretic assumptions about the philosophical significance of ACT details. Thus, any reading of the data that purports to discover what most people “naturally” believe can only telegraph what a particular researcher already believes. As such, any claim about what people really believe will rest on assumptions about the philosophical significance of certain facts. But the fitness of these facts to the principles of normative ethics, and their relevance to the metaphysics of causation, is beyond the scope of behavioral science, which can only account for their bearing on moral cognition in practice. Therefore, the dichotomy between a purely descriptive, ‘psychological’ error theory of moral-agentive attribution, and an inherently evaluative, ‘philosophical’ error theory, is false.

This insight is crucial in defusing the questions raised by previous philosophers who have asserted that “Abstract + Concrete = Paradox” (Sinnott-Armstrong, 2008). The variance in ordinary attributions of moral agency across contexts is not indicative of any inherent contradiction or paradox in ordinary beliefs about moral agency and determinism. In the experiment discussed here, there was no direct effect of ACT on FW. This revelation implies that there is nothing special or mysterious about “abstract framing” or “affective vignettes” that causes different behaviors undertaken in the same deterministic circumstances to be more or less frequently viewed as freely chosen.
4.7.4. Why do we tend to view others as autonomous agents in proportion to the extent to which they have ‘done wrong’?

Ordinary judgments of free will vary across moral contexts differ precisely because the philosophical concept of free will is derived from that of moral responsibility. Admittedly, the skeptic might be justified in taking the fickleness of ordinary moral-agentive attributions to lack philosophical significance. The problem with such a flippantly deflationist interpretation of the data presented here, however, is that these data do not speak to any such unreliability. To the contrary, the experiment discussed herein revealed attributions of free will to be highly reliable, and extremely predictable as an effect of moral responsibility attribution. If deflationists believe ordinary philosophical judgments to be so wildly unpredictable, so uninformed, and so unfounded, then they should be absolutely baffled by the regularity with which I have found ordinary people to use moral-agentive concepts. When statistically modeled as a moral construct arising from the notion of moral responsibility, there was nothing significantly unreliable about ordinary views regarding free will, and its compatibility with determinism. The consistency with which moral responsibility guides judgments of free will across contexts is striking, and its implications for analytic philosophy should not be overlooked and cannot be dismissed.

Philosophers may claim that what people believe about moral agency is none of their business, as philosophers. Nonetheless, how people come to have these philosophical beliefs (moral epistemology), whether they should act upon these beliefs (normative ethics), and why these beliefs turn out to be true or false (moral metaphysics) constitute three of the core problems of ethics and metaethics. Identifying these fault lines in the bedrock of philosophical conviction,
and studying the theoretical rifts and argumentative forces that arise from these conceptual divisions, are at the heart of analytic philosophy.

4.7.5. Why do philosophers care about free will?

Most philosophers seem to care about free will because of “the internal conception of agency and its special connection with the moral attitudes as opposed to other types of value” (Nagel, 1979). But postulating entities solely on the basis of their conformity to our internal conception of phenomena in the external world is not especially instructive or practical. Furthermore, philosophy stands to gain little by characterizing ordinary philosophical judgments as paradoxical, errant, unnatural, unreliable, or irrelevant. I therefore think that we should shift our efforts away from such causes, and focus instead on answering two more basic questions.

First, we might ask what guides our conceptual behavior in the moral-agentive domain. The past decade of research in moral psychology and experimental philosophy has already begun answering this question, and has yielded a wealth of knowledge to which I am indebted for many of the ideas I have discussed here. Second, we might ask what the psychological structure of moral-agentive constructs can reveal to us about debates in philosophy. The agency-last paradigm equips us for both lines of inquiry, providing a propitious opportunity for philosophical prospecting. On this approach, ostensibly competing ordinary attributions of free will might be brought into harmony. Might competing philosophical theories, which have heretofore been considered by many to be equally motley, be seen to have much more in common with one another when viewed through the lens of the agency-last paradigm? In the next (and final) section of this chapter, I offer an affirmative answer to this question.
There are several directions in which one might take this line of inquiry. A historian of philosophy might ask, as Nietzsche did, where “the thought that ‘the criminal deserves punishment, because he could have acted otherwise’” first arose within “the psychology of mankind in its early stages” (1887/1967). A cognitive psychologist might seek to study the component processes that lead to the ascription of certain moral-agentive attributes under various conditions. I think both of these approaches may be rife for innovative and fruitful projects. But in the space that remains, I only want to explain how the discovery of the psychological tendency to infer free will from moral responsibility may help us enrich our understanding of some of the most important insights in 20th century analytic philosophy. And so after disowning the idea that it is worth pursuing any error theory at all, and rejecting both the normative and deflationist stances, I expand in the final section on the deeper philosophical implications of the agency-last paradigm.

4.8. Agency-Last Perspectives on Classic Metaethical Puzzles

It may seem as though I have impugned some of the most fundamental beliefs of philosophers and non-philosophers alike, but this is not the case at all. I do not think that my beliefs are fundamentally all that different from those of most other philosophers. I am of the deepest conviction that in a literal sense, many of the most prominent figures in 20th and 21st century philosophy have tacitly operated within the agency-last paradigm. The views of many others can also be much better understood if we adopt this perspective. Just as the logically deduced taxonomies worked out by philosophers like Fischer (1987) have had major implications for the empirical study of moral cognition, discoveries in moral psychology—including those presented
Factlessness & Faultlessness

here—might help us better understand the philosophy of moral agency. The *agency-last paradigm* provides a theoretical and methodological instrument that not only helps us understand ordinary moral-agentive cognition and attribution, but which also yields a plausible philosophical framework that reduces free will to little more than a morally motivated accusation.

4.8.1. Contempt, Coercion, and Control

The *agency-last paradigm* can help us better understand Strawson’s claims in *Freedom and Resentment*, where he suggests that free will is only an account of others’ “attitudes towards us of goodwill, affection, or esteem on the one hand or contempt, indifference, or malevolence on the other” (1963). But attitudes like goodwill and contempt are inherently moralistic, and as Strawson himself points out, we can only hold these attitudes toward those we view as persons responsible for their own behavior. We could not, for instance, be contemptuous of a chalkboard for being the teacher’s pet. All of this suggests that asking whether an agent who is thought to be morally responsible does or does not have free will might only serve to beg the question, and that Strawson does not push his view far enough. Whereas Strawson suggests that free will may be fundamentally ineluctable, I see no reason why this should be the case. I think that questions about moral responsibility are primary to, and cannot be furthered by, inquiries into free will. The attitudes for which Strawson takes free will to be a kind of shorthand may or may not be held by those to whom we attribute them. But the automaticity with which we deride wrongdoers for their calculated willfulness seems to unveil ‘free will’ as nothing more than a moral epithet.

Like Strawson, Frankfurt recognizes the immediacy of moral responsibility from action. In *Alternate Possibilities and Moral Responsibility*, he describes a man who, in accord with his
own motivations, performs a certain action that another man would have forced him to perform, had he himself lacked the initiative to do so autonomously. In unpacking this case, Frankfurt points out that in at least some instances, the moral responsibility of an agent is decidable even if it remains unknown whether that agent could have freely chosen from among alternative possible courses of action:

The fact that a person was coerced to act as he did may entail both that he could not have done otherwise and that he bears no moral responsibility for his action. But his lack of moral responsibility is not entailed by his having been unable to do otherwise (1969).

But in trying to leverage free will into the space that Frankfurt’s moral argument opens up, his interpreters\(^2\) end up sapping free will of its primary value—subservience to the notion of moral responsibility.

\(^2\) I do not think it would be fair to attribute this interpretation to Frankfurt himself. My reading, though it seems to be an unpopular one, is that Frankfurt makes a conspicuous effort to establish moral responsibility without reference to the spooky notion of free will:

The two main concepts employed in the principle of alternate possibilities are "morally responsible" and "could have done otherwise." To discuss the principle without analyzing either of these may well seem like an attempt at piracy. The reader should rake notice that my Jolly Roger is now unfurled (1969).
The problems identified by Nagel in *Moral Luck* also largely resolve themselves if we reject the *agency-first* assumption, and instead take an *agency-last* approach. Nagel points out the (fairly obvious) fact that moral responsibility for what one has done depends on what one has actually done, and therefore also the circumstances in which one has done it and the outcomes to which it led. He recognizes that “from the point of view which makes responsibility dependent on control, all this seems absurd” (1979), but he also recognizes that the dependence of moral judgment on circumstance and outcome is clearly not absurd. In fact, its *independence* would be absurd, potentially leading to consequences like the jailing of all people who, while dialing their cell phones at the wheel of their cars, had the good fortune *not* to inadvertently strike and kill children. Nagel is deeply perplexed by the apparent tension between the principle that we are not responsible for outcomes outside of our control, and the realization that in real-life cases people are never or almost never in control of the outcomes of their behavior. This tension perplexes him so much that he describes it as paradoxical. But for whatever reason, he fails to recognize that the rejection of this paradoxical ‘Principle of Control’ is not the only way to dissolve the absurdity he discusses. Instead, one might simply reject the *agency-first paradigm*, which is at the root of his “point of view which makes responsibility dependent on control.”

4.8.2. The Puzzle of Moral-Agentive Cognition

Perhaps the most ironic place in which a potential *agency-last* solution to—or rather, dissolution of—the “abstract/concrete paradox” in moral psychology is approached but never quite captured is the very work that first brought that puzzle into the spotlight. In a footnote in that article, Nichols and Knobe acknowledge that “one might maintain that determinism is compatible with
moral responsibility but not with free will” (2007). This rather deft observation—which they credit to Fischer (1987), but whose philosophical roots, it can now be seen, run much deeper than that—is tantamount to an admission that factors that impinge upon free will do not necessarily impinge upon moral responsibility.

At the same time, Nichols seems to recognize that there is widespread acceptance that facts about $MR$ necessarily entail ones about $FW$. When *Science* published his repackaging of this same dataset from Nichols and Knobe (2007) as “The Experimental Philosophy of Free Will,” nobody blinked an eye at the fact that the data only reflected judgments of whether agents were “fully morally responsible” (2011). But it should be clear to the reader by now that using ‘free will’ as a stand-in for ‘moral responsibility’ in this way is deeply problematic. The presumable justification for taking such liberties is that one need not hold any particular position about free will in order to hold a given position toward moral responsibility, but that views on moral responsibility necessarily inform free will attitudes. But this, of course, is an agency-last justification. Thus, it still fails to answer the question of how research into free will matters, even if a deeper understanding of moral responsibility does. This, in turn, raises the question of what to make of the concept of free will itself. I conclude with my views on this subject.

4.8.3. Free Will as Moral Ether

Freedom of choice, unlike violations of moral principles, is not something that others *commit* and that we *observe*. Freedom of choice is only something that other people are said to *possess*. In this way, the concept of free will is by its very design opaque. Whereas the observation of transgressions and the assignment of responsibility for those transgressions are both significant
because of the causal roles they plays in individual and social behavior and regulation, free will occupies, at best, a latecomer’s role in social explanation. But the agency-last structure of moral-agentive cognition calls even this minimal value into question.

The notion of free will may have been central to a framework that people once had to assume in order to most successfully theorize about the connection between moral-agentive relata. In this sense, there may have been a time when appeal to free will could be justified as a (meta-theoretically desirable) form of inference to the best explanation. Over time, the centrality of that notion has been codified to the point of dogma, but the pre-scientific concept of free will has become so disconnected from the rest of our contemporary explanatory framework as to be merely nominal.

From an ontological perspective, the proposed existence of free will is not so much false, as it is an empty hypothesis. Moral responsibility without free will is only inconceivable in the trivial sense in which “space without ether is unthinkable” (Einstein, 2010/1920).” In this sense, free will is only a kind of moral ether.
CHAPTER FIVE

ADJUDICATING ADJUDICATION

In philosophy we do not draw conclusions. "But it must be like this!" is not a philosophical proposition. Philosophy only states what everyone admits.

-Ludwig Wittgenstein, “Philosophical Investigations”

Does evidence of a biomechanical cause of psychopathy reduce sentencing to the same extent for male and female judges? An experiment published last year found that when psychiatric evidence of criminal psychopathy was supplemented by evidence of an underlying biomechanism, judges assigned shorter average sentences and were more likely to cite at least one mitigating factor of psychopathy in accompanying written opinions. But it remains unclear whether the absence of neurobiological evidence justifies the retention of longer sentences, and unclear whether the opinions of this judicial sample are widely held, or reflect the unique demographics of the U.S. state trial judiciary. Here, a reanalysis of the data in which this effect was first revealed found no effect of biomechanism on female judges’ sentencing or written opinions. These results suggest that it is worth further investigating whether the overrepresentation of men on the bench may lead to a hard-scientific bias in U.S. state courts. Additionally, the findings highlight the need to develop a scientific understanding of the social forces that give rise to these gender differences in the first place.
5.1. The Double-Edged Sword

As scientific understanding of the relationship between neurobiology and behavior becomes increasingly sophisticated, policymakers are faced with the challenge of assuring that this new knowledge is applied in ethical ways. There seems to be implicit agreement among scientists and philosophers that it is our responsibility to adhere to a *principle of epistemic caution*, according to which our neuroethical policies should never outpace our neuroscientific knowledge. But this principle may have the paradoxical effect of preventing some life-or-death decisions from being made on the basis of uncertain scientific premises, only to allow those choices to be decided by systemic social inequalities. Here, I discuss a neurogenetic allele whose influence on behavior is unclear to scientists, whose potential philosophical implications remain disputed by ethicists, yet whose exclusion from the courtroom has clear and troubling inegalitarian implications.

In a recent experiment, Aspinwall, Brown, and Tabery (2012a) found that U.S. state trial judges assigned shorter average sentences to criminal psychopaths when psychiatric diagnoses were accompanied by supplementary evidence of an underlying neurobiological cause, rather than being presented without biomechanical explanation. Differences in sentencing seemed deliberative, as supplementary neuroscience testimony led to more frequent mention of mitigating aspects of psychopathy in judges’ written explanations of their reasoning, and to decreased Likert ratings of the extent to which the evidence of psychopathy aggravated punishment. These findings highlight several issues in determining the appropriateness and admissibility of neuroscience testimony in the courtroom.
5.2. Scientific & Philosophical Skepticism

It remains unclear to neuroscientists how or even whether biomechanisms like the one described to the judges in this experiment relate to recidivism (Tikkanen et al., 2011) and reform (Lester & Eley, 2013). The psychiatric diagnosis provided to judges in both experimental conditions (Hare et al., 1990) is actually a much better predictor of future violence than the neurogenetic polymorphism whose presence or absence made such a difference in sentencing. In fact, the same allele thought to put those who have not experienced childhood maltreatment at increased risk for violent behavior appears to place those who were not mistreated as children at decreased risk for that same kind of behavior (Tabery, 2009).

There are also philosophical reasons to doubt that there is anything especially exculpating about the loss of control to biomechanical factors, as opposed to more ordinary factors like poverty, childhood maltreatment, and dumb luck (Nagel, 1979). Many bioethicists reject the idea that there is any “bright line” distinguishing the appropriate consideration of genetic and non-genetic influences on behavior (Silvers and Stein, 2003). All of these considerations suggest that the sentences judges assign to criminal psychopaths may be biased by the presence of neuroscience evidence whose forensic and moral value is uncertain. Nonetheless, there are reasons to think that the absence of this same type of evidence might bias the judicial system toward the opinions of male judges.

Previous research has found that women tend to be more skeptical of biotechnologies than men tend to be, especially when policies rely on the opinions of experts, rather than those of the public (Simon, 2010). One study found that scientific knowledge led women to be more skeptical toward scientific experts, whereas it led men to be more credulous toward experts
(Hayes & Tariq, 2000). These gender differences in attitude toward science, together with the seductive allure of neuroscience explanations (Weisberg et al., 2008), raise an important question: might the hard-scientific bias seen among U.S. state trial judges be an artifact of the gender imbalance in the American judiciary?

5.3. Experiment

Here, I present a reanalysis of the data from Aspinwall et al. (2012a, 2012b), which was conducted in order to test the hypothesis that gender moderated the effect of biomechanism on judges’ sentencing of criminal psychopaths. The results suggest a clear-cut matter of injustice: only male judges, who account for 73% of U.S. state trial judges (Kimball et al., 2013), assigned significantly longer sentences ($M = 3.556$ years, $SEM = 1.498$) and were significantly less likely to cite evidence of psychopathy as a mitigating factor ($OR = 0.3671$, 95% CI [0.1793, 0.7517]) when biomechanical evidence was absent (Figure 12).
**Figure 12.** Influence of Additional Biomechanism on Sentencing by Male and Female Judges

*Figure 12.* $N$ indicates the total number of judges of each gender in each experimental condition, collapsed across presenting party (1,3). (A) Average sentencing with and without evidence of low MAOA activity, by gender of judge. Based on estimated marginal means from ANOVA, collapsed across presenting party (1,3). Error bars represent ±1 SEM. (B) Percentage of judges who mentioned at least one mitigating factor concerning the evidence of psychopathy in their written explanations of the reasoning behind the sentences they rendered.

### 5.4. Discussion

These data suggest a male judicial bias in favor of unverified ‘hard science’ (Haberstick et al., 2014) over better-verified ‘soft science’ (Grann et al., 1999). Lacunae in knowledge always necessitate a reliance on opinion, and there is nothing inherently unethical or unscientific about making value judgments in interpreting contested scientific claims. But allowing the gender composition of the judiciary to decide between competing systems of scientific values—the value systems according to which ambiguous neurobiological testimony is interpreted—is both unscientific and unfair.
The limited experimental sample of female judges leaves open the possibility of a small but undetected effect of biomechanism on the sentences they assigned. Still, the results imply that the appropriate treatment of neuroscience evidence in the courtroom does not just depend on the consequences of its admission, but also on the consequences of its inadvertent omission and its deliberate exclusion. This means that we cannot wait until neuroscience advances before we address potential biases in neuroethical policy. Decisions about the admissibility of this kind of neuroscience testimony are already being made (Forzano et al., 2010), and to the extent that these decisions might increase reliance on inegalitarian values, their biasing effects must be adjudicated now. But how?

The first step in addressing the potential over-representation of a hard-scientific bias among male judges is to raise awareness of the issue. In deciding how to act on this knowledge, it may be essential to recognize that despite its scientific merits, the principle of epistemic caution might itself need to be applied with caution in cases that are entangled with issues of social justice, such as gender inequality on the bench. Moving forward, it may also prove invaluable to study the psychological and sociological factors that underwrite these gender differences in judicial opinion, and gender differences in public attitudes toward science more generally. Doing so might help us not only to identify and remedy their effects on judicial sentencing, but also to understand and address the social forces that give rise to these gender differences in the first place.
Returning to the focal question of this project, we can once again ask: why do people disagree about the answers to philosophical questions? Many answers have been proposed, and my suggestion has been that many answers are, in fact, correct in different cases. It all depends on what kind of philosophical question is being asked. Furthermore, if the causes of philosophical dispute are diverse, then the traditional analytic method will only work for some disputes, and we will need other methods to work with disputes whose underlying causes are different from those usually anticipated by analytic philosophers. This is why it may be so valuable to have the taxonomy that lets us understand the type of issues at stake. Once we know what kind of disagreement something is, we can better draw on our understanding of that type of disagreement to better develop and test hypotheses about how to resolve it.

6.1. Recommendations

In order to adjudicate philosophical disputes, interlocutors need to ask each other the right questions, in order to establish which issues are at stake, and which are being set aside (at least for the purposes of a given dispute). To bring disputants together in agreement over a single conclusion, we first need to identify where party lines are being drawn. If intuitions were always truth-evaluable—that is, if every intuition were about facts that were either true or false—then philosophers would be justified in using contradiction and negation to infer from true intuitive premises that all conflicting intuitions are false. But I have argued the intentional object (Searle,
Factlessness & Faultlessness

Holtzman 104

1980) of an intuition very often lacks facticity, because intuitions depend on factors that naturally vary between individuals.

A dispute that is ostensibly about some object or phenomenon in the external world cannot be adjudicated if that dispute arises, at least in part, because the object or phenomenon at issue for one disputant is not the same as that for another disputant. As we saw in Chapter Two, because a robot’s ability to undergo the phenomenology of love leaves ambiguous what is meant by love itself, the phenomenology being debated will differ from person to person. It follows from this that the truth of a given evidential foundation—and in particular, the propriety of some intuition—does not necessarily imply that intuitive evidence in conflict with it is false or improper in some way.

The fact that one disputant is right to assert something does not mean that her interlocutor would necessarily be wrong to deny it. Therefore, philosophical insights are not necessarily ones that tell us one theorist is right and her detractor is wrong. Instead, we will often gain insight by understanding how the theorist and her detractor could have both come to hold reasonable, accurate views, even though those views appear to stand in contrast to one another. The study of individual differences is crucial to a strong understanding of why people disagree about the answers to philosophical questions, and this understanding is essential to understanding what the very topic or topics at stake in a philosophical dispute are.

The traditional approach to resolving philosophical disputes also has a decidedly normative dimension. Only one intuition can be right, and all others are wrong; only one tree of deductive inferences from that intuition is good, and everything else is bad. This disease of equating description (of arguments) with normative claims about those arguments is, and has for a long time been, one of the main diseases of philosophy. My proposal has been to respond to
these considerations by deemphasizing the question of what people *should* believe. Instead, it is more informative and less presumptuous to focus on *why* and *in what sense* they believe it.

The taxonomy I have proposed helps focus questions in these domains. People may hold conflicting views because at least one of them is reasoning incorrectly, or because there are multiple rationally acceptable modes of inference about a question—because the dispute is *faultless*. And people may hold conflicting views because the sense in which they mean whatever it is they claim to believe differs—that is, because the putative concepts being referred to by disputants, and the facts in dispute, are actually slightly different from one another.

### 6.2. Shortcomings & Future Directions

The presentation herein has been taxonomic in nature, and while I have argued that the development and defense of a taxonomy of philosophical dispute is the central value of the work as it is presented here, the use of these data in that mode also has major downsides. While I have striven for cohesion of the case study chapter structure to the taxonomy outlined in Chapter One, perhaps my biggest disappointment in the foregoing chapters is their lack of cohesion to one another. However, this disappointment needs to be qualified in two ways.

First, I wish to make clear the project you see in front of you is incomplete. Filling in the remaining interstitial spaces with the missing pieces of theory and evidence is a project that I have set out for myself as a long-term research program, and will require more empirical and literature-based research. Second, I wish to make clear that the lack of cohesion is in large part a shortcoming of the chapters and their portrayals of the research projects, not a problem of the research projects themselves. The theoretic motivations for each study were drawn from the
studies that came before it. But so far, the data I have collected is only consistent with—not evidence for—the hypothesis that the underlying cause of gender differences in all three studies is related.

These chapters represent the progression of a single project over several years. Chapters Two and Three fit rather neatly into the traditional philosophical categories of subjectivism or relativism (in the case of Chapter Two), and realism or objectivism (in the case of Chapter Three). The purpose of this project has been to get away from this facile dichotomy. Additionally, each chapter was, at some point, a “future direction” suggested by the last. For these reasons, I focus this discussion of future directions on Chapters Three and Four, both of which I am in the process of expanding. Their expansion represents two distinct lines of research—one in moral psychology, and one in neuroethics—but as Chapter Four was initially a “future direction” suggested by Chapter Three, these lines of research continue to influence one another.

6.2.1. Expanding the Agency-Last Paradigm

After noticing that people tend to employ different responsibility constructs when asked to place blame in different situations, we might ask what guides this behavior. Elsewhere, I argue that people do not tend to toggle between different constructs of moral responsibility on the basis of the situations that they encounter. Instead, I suggest that those who endorse a definition of responsibility-as-accountability tend to place blame in proportion to the severity of moral transgressions and actual involvement in those transgressions, whereas those who endorse a view of responsibility-as-capacity (most commonly, the capacity to do otherwise) are largely
indifferent to this sort of information. Second, we might ask what this new, non-error-theoretic taxonomy might reveal about the debate between compatibilists and incompatibilists, especially when examined through the lens of the agency-last paradigm. However, this question could not be studied in sufficient depth without running experiments outside the scope of the philosophical work you seen in front of you.

Evidence collected for Chapter Four supports the alternative hypothesis that instead, responding differently to our different cases can be accounted for by the notion of moral responsibility people endorse in the abstract. These data were collected, like the rest of the data presented here, with survey methods. This method has a number of limitations, many of which have been discussed here. But most relevant to the expansion of the agency-last paradigm, paper-and-pencil surveys that ask directly about moral responsibility in a concrete case cannot also be used to collect data about how much, abstractly speaking, an individual thinks that the nature of a crime bears on moral responsibility for that crime.

6.2.2. Better Adjudicating Adjudication

At the end of Chapter Five, I stop short of making strong, concrete recommendations for addressing the gender discrepancy in the sentencing of criminal psychopaths. I do this on principle, because of a lack of clarity in the data discussed in that chapter. It remains unclear whether the gender differences described in that chapter are due to a bias among male judges for of unverified ‘hard science’ (a form of scientism), a bias among female judges against unverified ‘hard science’ (a form of skepticism), or some combination thereof.
Furthermore, there is an even more basic characterization nature of the differences underwriting this bias cannot be inferred from this experiment. It cannot even be said with certainty whether these gender differences in belief are moral or scientific in nature. Presently, I am working a project aimed at untangle the many logical and argumentative kinks in the data discussed in that chapter. Although I hope that I (and others) may engage in further empirical research into the causes of these gender differences and potential ways to adjudicate experimental data with only yield theoretic clarity after this untangling is complete. There are a number of confounded issues in the data that need redress, but I think that the keystone issue is identifying which of the following theses manifests the gender differences underwriting the discrepancy in sentencing mitigation:

1. Biomechanism warrants a reduction in sentencing.
2. Psychopathy warrants a reduction in sentencing.
3. Biomechanism is evidence of psychopathy.

Thesis (1) is a view about the presence of external influences on an agent’s behavior, and their bearing on sentencing. In contrast, (2) regards the absence of internal influences that might bear on sentencing. Proposition (3) differs fundamentally from both (1) and (2), as it regards evidence for scientific diagnosis, rather than grounds for judicial prescription. The distinction between (1) and (2) can be further illustrated by attending to differences in the actual testimony read by judges in each experimental condition. In the neurobiologist’s testimony, blame is offloaded onto genes through a clear-cut domino effect, whereby genes cause a broken
mechanism, a broken mechanism causes an inability to learn, and an inability to learn prevents healthy socialization:

Psychopaths, because of their genetically-induced dysfunctional violence-inhibition mechanism, do not learn to associate distress in others with anxiety in themselves and are thus resistant to moral socialization.

Notably, the psychiatrist’s testimony does not provide a less scientific-sounding explanation of why the criminal did not socialize like the rest of his peers. It provides no explanation at all, instead only begging the question why the criminal is so poorly-socialized in the first place: “psychopaths are resistant to moral socialization because of their disorder,” i.e. psychopathy. Thus, it remains unclear whether male and female judges differed in their treatment of neurobiological versus psychiatric testimony, or in their requirements for causal explanations of moral patiency.

6.3. Conclusion

Neither of these future directions represents the first time I have suggested that, first and foremost, we need to address issues of conceptual disentanglement. The prime importance of that goal was hypothesized at the beginning of the opening chapter, in which I suggested that only by identifying the contours of philosophical disputes, and the factors that draw people across those fault lines, can we even begin to really understand the concepts and theories in dispute.
Notably, because turns out to be a plurality of causes for the diversity of philosophical belief—and here, I have identified just a few among many—then no single method of inquiry or dialogue is sufficient for scrutinizing the entirety of the philosophical landscape. Instead, a variety of approaches is called for. But in order to know what approach to take to a given philosophical question—that is, what kind of answer is even to be sought—we need to employ the taxonomy of causes of philosophical dispute that I have outlined and illustrated in the preceding chapters. Such a taxonomy has helped us better identify the causes of the philosophical disputes discussed in this work, and can help us to do so in many more cases, as well as help us discover potential modes of resolution to such disputes. It is for this reason that I have used this forum to develop, defend, and demonstrate a taxonomy of philosophical dispute that covers not only the rational-irrational and true-false dimensions of dispute, but the faulty-faultless and factive-factless dimensions as well.
1. Suppose scientists figure out the exact state of the universe during the big bang, and figure out all the laws of physics as well. They put this information into a computer, and the computer perfectly predicts everything that has ever happened. In other words, they prove that everything that happens, has to happen exactly that way because of the laws of physics and everything that's come before. In this case, is a person free to choose whether or not to murder someone?

2. Suppose you drive to the local baseball stadium with some friends, and try to buy tickets at the door. There are 7 of you, but there are only 6 tickets left. You can either drive everyone to a nearby bar, which will be a lot less fun than being at the game, or 6 of you can go in, and 1 of you can take the bus home and miss the game entirely. Is it most fair for everyone to go to the bar?

3. Suppose a mad scientist takes out your brain, and puts it in your best friend’s head. During the same operation, the scientist takes out your friend’s brain, and puts it in your head. Now your body has your friend’s brain, and your friend’s body has your brain. Your heroic mother storms into the room to save you, but not your friend, who she believes got you into this mess. Is the person with your body still you, her son?

4. Suppose neuroscientists are able to identify every part and every connection in the human brain. Working with a team of computer scientists, they then build a robot that has a complete electronic replica of the human brain. Could this robot experience love?

5. Suppose that all you know about Einstein is that he developed the Theory of Relativity. But suppose it turns out that Einstein actually stole the idea from some guy named Moynahan,
who nobody has ever heard of. In this case, when you use the name “Einstein,” are you actually referring to Moynahan?

6. Suppose you hear the sound of your cell phone, so you reach in your pocket and answer the call. Your landlord is on the line, but you realize later that your ringer was off, and the sound you heard was actually someone else’s phone. When you heard that other person’s phone ring and mistook it as your own, did you actually know someone was calling you?

7. Suppose you meet a man from the future who knows everything there is to know about science. He tells you that he doesn’t like apples, and says that though he has never eaten one, he has figured out what apples taste like just by studying the relevant science. Could he know what apples taste like without ever having eaten one?

8. Suppose scientists are able to use stem cells to grow lungs that breathe without being connected to a body. They then grow a heart that pumps without being connected to a body. If they can do all this, can they create a brain that thinks without being connected to a body?

9. Suppose a runaway train is coming down a track, and is certain to kill five workmen who can't get out of the way. You're standing next to the controls and can switch the train to the other track, but if you flip the switch, one man working on that track is sure to die. Should you flip the switch?
References

Adleberg, T., Thompson, M., & Nahmias, E. (2013). Do women and men have different philosophical intuitions? A failure to replicate results from Buckwalter & Stich.


Frege, G. (1879). Begriffsschrift, a formula language, modeled upon that of arithmetic,


Factlessness & Faultlessness

New York: Guilford Press.


Holtzman, G. S. (under review A). The agency-last paradigm: Free will as moral ether.

Holtzman, G. S. (under review B). Adjudicating Adjudication: Gender Moderates the Effect of Biomechanism on Judges’ Sentencing of Psychopaths


Factlessness & Faultlessness

Studies, 151(2), 299-327.


Nichols, S. (2011). Experimental philosophy and the problem of free will. Science, 331(6023),


