Oxford Studies in Experimental Philosophy
# Contents

## Introduction

Tania Lombrozo, Joshua Knobe, and Shaun Nichols  
1

## 1. Huck vs. JoJo: Moral Ignorance and the (A)symmetry of Praise and Blame

David Faraci and David Shoemaker  
7

## 2. The Cognitive Mechanisms of Intolerance: Do our Meta-Ethical Commitments Matter?

Jennifer Cole Wright, Cullen B. McWhite, and Piper T. Grandjean  
28

## 3. Doing, Allowing, and Enabling Harm: An Empirical Investigation

Christian Barry, Matthew Lindauer, and Gerhard Øverland  
62

## 4. Attributing Responsibility: Actual and Counterfactual Worlds

Tobias Gerstenberg and David A. Lagnado  
91

## 5. The Moral Status of an Action Influences its Perceived Intentional Status in Adolescents with Psychopathic Traits

Elise M. Cardinale, Elizabeth C. Finger, Julia C. Schechter, Ilana T.N. Jurkowitz, R.J.R. Blair, and Abigail A. Marsh  
131

## 6. The Concept of Intentional Action in High-Functioning Autism

Edouard Machery and Tiziana Zalla  
152

## 7. A Scientific Case for Conceptual Dualism: The Problem of Consciousness and the Opposing Domains Hypothesis

Anthony I. Jack  
173

## 8. Carving Up the Social World with Generics

Sarah-Jane Leslie  
208
   Deborah Tollefsen, Roger Kreuz, and Rick Dale

10. The Good in Happiness 253
    Jonathan Phillips, Sven Nyholm, and Shen-yi Liao

11. Causal Reasoning: Philosophy and Experiment 294
    James Woodward

    Taylor Shaw Murphy

Index 369
Introduction

Tania Lombrozo, University of California, Berkeley
Joshua Knobe, Yale University
Shaun Nichols, University of Arizona

The aim of Oxford Studies in Experimental Philosophy is to provide a venue for publishing work in this new field. Experimental philosophy has emerged over the last decade as the methods of psychological science have been brought to bear on traditional philosophical problems. The ambition of the field has been to use these empirical methods as a way to illuminate some of the oldest issues in philosophy.

Experimental philosophy is an interdisciplinary field at the intersection of philosophy and psychology. Accordingly, just about all of the papers in this field so far have been published either in philosophy journals or in psychology journals. There is certainly value in this approach, especially insofar as it has led to real engagement with work in each of these disciplines, considered separately. However, there is also a downside. When people publish in philosophy journals, they may feel a certain pressure to focus exclusively on the philosophical side of their work, to the exclusion of the more psychological side; there may be a corresponding pressure to focus on the empirical when publishing in psychology.

OSEP is intended to serve as a venue in which people can explore experimental philosophy in all of its aspects, addressing both philosophical and psychological issues, as needed. Looking through this first volume, one finds a mix of chapters by philosophers, chapters by psychologists, and chapters co-authored by people in both disciplines. Yet the work done by these different people does not seem to be easily classifiable
into the different disciplines. Instead, one finds the emergence of a truly interdisciplinary field in which people from different disciplines are working together to address a shared set of questions.

The volume is roughly structured into four sections. The first three chapters focus on recent developments in moral psychology—a topic that has seen lively debate and a great deal of progress over the last decade. In the second section, we highlight three contributions that bring new methods to moral psychology: formal modeling and special populations. In the third section, we assemble four chapters that adopt an experimental philosophy approach to novel topics, including intuitive dualism, generics, joint action, and happiness. Finally, in the fourth section, we present two chapters that provide critical and historical context to the development of experimental philosophy.

**Moral Psychology: New Findings**

David Faraci and David Shoemaker investigate the *prima facie* plausible idea that our judgments of blame and praise should be *symmetrical*. For instance, if the unfortunate circumstances of a person’s upbringing render him less blameworthy for bad actions, it should also render him less praiseworthy for good actions. However, in a series of experiments, they find that people tend to treat such cases asymmetrically. At least in some cases, a bad upbringing makes a person less blameworthy for bad actions but *more* praiseworthy for good actions. Faraci and Shoemaker explore various ways in which one might square these results with the idea that blame and praise should be treated symmetrically.

Jennifer Cole Wright and colleagues examine the relationship between two ethical issues that might at first seem quite distinct: (a) the meta-ethical question as to whether certain moral claims are objectively true and (b) people’s willingness to show toleration for those who have opposing values. Wright and McWhite find that their experimental participants tended to be “meta-ethical pluralists,” meaning that each participant tended to think that some moral questions have objective answers while others do not. They then showed a surprising relationship between issues (a) and (b). Specifically, there was a significant correlation whereby participants showed less tolerance for people with opposing values about those questions they regarded as being objective.
Christian Barry and colleagues challenge a long-standing distinction between doing and allowing harm. Backed by a series of novel experimental findings, the authors’ suggestion is that this bipartite distinction has neglected an important third category: enabling harm. The authors present evidence that judgments about harm can reliably be classified as doing, allowing, or enabling harm, and that this classification has consequences for people’s judgments of moral responsibility. Differentiating “enabled harm” from other cases provides a more fine-grained taxonomy of harm that can inform both normative and empirical work in moral psychology.

Moral Psychology: New Methods

Tobias Gerstenberg and David Lagnado consider how moral responsibility is allocated for outcomes that result from the actions of multiple agents. According to most accounts of responsibility attribution, agents should be held responsible to the extent that their actions made a difference to the outcome. Yet in some cases we hold agents responsible for outcomes even when those outcomes do not depend (counterfactually) on their individual actions. For example, we might hold a shooter responsible for the death of a victim even if a second shooter fired simultaneously, and that shooter’s actions were sufficient to bring about the death. Gerstenberg and Lagnado suggest that we can make sense of such cases by considering not only whether an agent’s actions made a difference in the actual situation being considered, but also whether they would have made a difference in other possible situations, such as those in which a second shooter was not present. The authors develop this idea in terms of counterfactuals defined over structured causal models and discuss both published and new results that support their formulation.

Judgments about whether an agent intentionally produced an outcome are affected by whether the outcome is morally bad or morally good. For instance, if a person performs an action that he knows will harm the environment, then even if he did not intend to harm the environment people tend to say that he intentionally harmed the environment. Elise Cardinale and colleagues investigate whether this phenomenon occurs in adolescents with psychopathic traits. This population has known deficits in moral cognition, and it has been quite unclear whether these deficits in
moral cognition would lead to atypical performance concerning intentionality. Cardinale and colleagues found no difference between the adolescents with and without psychopathic traits—an interesting result with implications for current accounts of moral evaluation and intentionality attribution.

Edouard Machery and Tiziana Zalla explore a related question about the understanding of intentional action among participants with high-functioning autism. In earlier work, normally developing adults were presented with a person deciding to pay extra to get a benefit (a smoothie). As expected, the participants said that the person intentionally paid extra. By contrast, when presented with a case in which a free benefit—a cup—was given to the person, participants tended to deny that the person intentionally got the cup. Machery and Zalla presented this task to participants with high-functioning autism and found that these participants did not distinguish the cases. Machery and Zalla appeal to deficits in social cognition to explain this result. In particular, they suggest that people with high-functioning autism have difficulty recognizing that a person might have an instrumental desire for something that is intrinsically disvalued.

**New Topics in Experimental Philosophy**

A prominent theme in contemporary cognitive science is that people are dualists about mind and brain. At least in many contexts, people regard mind and brain as fundamentally independent kinds of things. Anthony Jack adduces evidence for this view and introduces a novel explanation for folk dualism. He argues that the cognitive tension between reasoning about the mind (“mentalizing”) and reasoning about physical mechanisms reflects a tension at the level of opposing neural systems. In particular, when one of these systems is activated, it actually suppresses activation of the opposing system.

Turning to an issue with more immediate practical implications, Sarah-Jane Leslie discusses a series of experimental studies on the use of generic language. Recent work in philosophy, psychology, and linguistics has explored in depth the semantics of generic sentences (e.g. “Tigers are striped.”). Leslie now asks what impact those sorts of sentences have when they are applied to social categories (“Women are . . .”, “African-American people are . . .”). She shows that these sentences
communicate a particular conception of social categories involving what psychologists call essentialism. She then argues that this essentialist conception is tied to a detrimental understanding of human abilities.

Deborah Tollefsen and colleagues use experimental methods to explore people’s intuitions about joint action. Philosophers distinguish things that human beings do together (having a conversation, singing a duet) from those that people do individually (reading a book, singing a solo). Tollefsen and colleagues use multidimensional scaling to examine the factors that make people regard various actions as similar or different. They obtain a two-dimensional solution, with one dimension corresponding surprisingly neatly to this philosophical notion of “jointness.” In other words, even when people are not being asked questions that explicitly involve the concept of jointness, their spontaneous judgments of similarity show a clear influence of that concept.

Jonathan Phillips and colleagues present a series of studies investigating people’s ordinary concept of happiness. The proposal that they develop and test combines elements from two traditions of research on happiness. Like philosophers endorsing descriptive accounts of happiness, as well as many psychologists engaged in the empirical study of happiness, they suggest that attributions of happiness are largely based on the psychological states of the person being considered, not on the normative status of the life that person actually led. Yet more in line with philosophers offering evaluative views of happiness, they also find that such judgments depend on whether people believe that the agent in question ought to have experienced a positive psychological state in a given case. The result is a novel hybrid of descriptive and normative theories of happiness attributions, with interesting implications for both philosophical and psychological work.

Critical and Historical Reflections on Experimental Philosophy

James Woodward offers an analysis and review of research on causal judgment, covering both philosophical work from a broadly interventionist perspective and empirical work from a variety of psychological disciplines including cognitive and developmental psychology. Woodward considers how normative accounts of causation have informed and
can inform empirical investigation; he also provides guidelines for the type of research most likely to have interesting philosophical implications. In particular, Woodward discusses research from both survey-based experimental philosophy and measures of causal learning and reasoning from psychology, offering both critical and constructive suggestions about their relative merits.

Taylor Shaw Murphy engages in a more historical investigation, exploring an important precursor of contemporary experimental philosophy in a movement that flourished from roughly 1935 to 1965. The experimentalists of this earlier period conducted empirical tests of some of the key theses coming out of ordinary language philosophy, and Murphy describes some of their most striking experimental results. In addition, he explores the metaphilosophical debate that emerged at that time, examining the many ways in which it prefigured the contemporary debate over experimental philosophy.

In sum, this volume highlights both the progress and the promise of experimental philosophy, bringing together contributions on established topics alongside interdisciplinary work that combines empirical methods with philosophical issues in new and exciting ways. Perhaps most notable of all, however, is that the integration of philosophy and psychology is so seamless, suggesting that experimental philosophy is truly coming of age.
1

Huck vs. JoJo

Moral Ignorance and the (A)symmetry of Praise and Blame

David Faraci, University of North Carolina
David Shoemaker, Tulane University

The more I studied about this the more my conscience went to grinding me, and the more wicked and low-down and ornery I got to feeling. And at last, when it hit me all of a sudden that here was the plain hand of Providence slapping me in the face and letting me know my wickedness was being watched all the time from up there in heaven, whilst I was stealing a poor old woman’s nigger that hadn’t ever done me no harm, and now was showing me there’s One that’s always on the lookout, and ain’t a-going to allow no such miserable doings to go only just so fur and no further, I most dropped in my tracks I was so scared. Well, I tried the best I could to kinder soften it up somehow for myself by saying I was brung up wicked, and so I warn’t so much to blame; but something inside of me kept saying, “There was the Sunday-school, you could a gone to it; and if you’d a done it they’d a learnt you there that people that acts as I’d been acting about that nigger goes to everlasting fire.” (31.19) It made me shiver. And I about made up my mind to pray, and see if I couldn’t try to quit being the kind of a boy I was and be better. So I kneeled down. But the words wouldn’t come. Why wouldn’t they? It warn’t no use to try and hide it from Him. Nor from ME, neither. I knowed very well why they wouldn’t come. It was because my heart warn’t right; it was because I warn’t square; it was because I was playing double. I was letting ON to give up sin, but away inside of me I was holding on to the biggest one of all. I was trying to make my mouth SAY I would do the right thing and the clean thing, and go and write to that nigger’s owner and tell
where he was; but deep down in me I knowed it was a lie, and He
knowed it. You can’t pray a lie—I found that out. (31.20) [After
writing the note,] I felt good and all washed clean of sin for the first
time I had ever felt so in my life, and I knowed I could pray now.
But I didn’t do it straight off, but laid the paper down and set there
thinking—thinking how good it was all this happened so, and how
near I come to being lost and going to hell. And went on thinking.
And got to thinking over our trip down the river; and I see Jim
before me all the time: in the day and in the night-time, sometimes
moonlight, sometimes storms, and we a-floating along, talking and
singing and laughing. But somehow I couldn’t seem to strike no
places to harden me against him, but only the other kind. I
’d see him standing my watch on top of his’n, ’stead of calling me, so
I could go on sleeping; and see him how glad he was when I come
back out of the fog; and when I come to him again in the swamp, up
there where the feud was; and such-like times; and would always
call me honey, and pet me and do everything he could think of for
me, and how good he always was; and at last I struck the time
I saved him by telling the men we had small-pox aboard, and he was
so grateful, and said I was the best friend old Jim ever had in the
world, and the ONLY one he’s got now; and then I happened to
look around and see that paper. (31.23) It was a close place. I took it
up, and held it in my hand. I was a-trembling, because I’d got to
decide, forever, betwixt two things, and I knowed it. I studied a
minute, sort of holding my breath, and then says to myself: “All
right, then, I’ll GO to hell”—and tore it up. (31.24, 31.25)

In these famous scenes from The Adventures of Huckleberry Finn, Huck
is trying desperately to decide what to do with Jim, the slave he has been
traveling with down the river, the man who is, to Huck’s mind, someone
else’s property. When given the opportunity to return Jim to his “rightful
owner,” Huck ultimately decides to go against his upbringing, his con-
science, and societal norms in not turning Jim in. Consequently, by his
own lights he is going to hell for doing what he sincerely believes is the
wrong thing. By our lights, though, he is clearly doing the right thing,
despite his morally deprived upbringing. Is he thus praiseworthy?

Compare the following case:

JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small,
undeveloped country. Because of his father’s special feelings for the boy, JoJo is
given a special education and is allowed to accompany his father and observe his
daily routine. In light of this treatment, it is not surprising that little JoJo takes his
father as a role model and develops values very much like Dad’s. As an adult, he
does many of the same sorts of things his father did, including sending people to
carceration or to death or to torture chambers on the basis of whim. He is not coerced
to do these things, he acts according to his own desires. Moreover, these are
desires he wholly wants to have. When he steps back and asks, “Do I really want
to be this sort of person?” his answer is resoundingly, “Yes,” for this way of life
expresses a crazy sort of power that forms part of his deepest ideal.¹

In torturing a peasant on a whim, JoJo goes along with his upbringing
and conscience, doing what he sincerely believes is the right thing but
what we know is the wrong thing. Is he thus blameworthy?

Both Huck and JoJo were raised in morally blinkered environments,
and they have both consequently come to accept deeply mistaken moral
views: Huck thinks he ought to return Jim to his “owner”; and JoJo thinks
it is morally permissible (or perhaps obligatory) to beat peasants when he
feels like it.

There have been (at least) two assumptions that theorists have thought
it natural to make in cases such as these. First, it seems plausible to some
that moral deprivation of this sort is a responsibility-undermining factor,
that those who were raised in such morally blinkered environments are
off the hook when it comes to assignments of moral responsibility. So
JoJo, we might think, is excused for beating the peasants given the extent
of his childhood moral deprivation. Call this the Moral Deprivation–
Excuse Thesis (the MDE Thesis).² Second, it is widely believed that
praiseworthiness is the positive analogue of blameworthiness, and thus
that factors affecting judgments of moral responsibility should have
symmetrical effects on judgments of praise- and blameworthiness. Call
this the Symmetry Thesis.³ On a natural interpretation of this thesis, if
poor moral upbringing mitigates assignments of JoJo’s blameworthiness,
it ought likewise to mitigate assignments of praiseworthiness in cases of
equivalent moral deprivation.

In this essay, we examine how certain sorts of moral knowledge depriva-
tions in an agent’s upbringing bear on people’s actual assessments of that

² This is the view taken by Wolf, as we will detail below.
³ The Symmetry Thesis is widespread. It is entailed by what Doris and Knobe call the
more general assumption of invariance in moral responsibility assessments. For discussion
and citation, see John Doris and Joshua Knobe, “Strawsonian Variations: Folk Morality and
(Oxford: Oxford University Press, 2010).
agent’s responsibility. At first blush, the data we have collected appears to cast doubt on both the MDE and Symmetry theses. First, our experimental results suggest that people do not, in fact, view deprivation as wholly morally excusing (though they do not view it as irrelevant, either). Second, our data appear to suggest that people have commitments to distinctly different conditions for blame- and praiseworthy agency. This seems to cast doubt on the Symmetry Thesis.

In both cases, however, we argue that the correct response to our data is not simply to reject the theses in question. First, our results suggest that the relationship between deprivation and moral assessment is more nuanced than straightforward acceptance or denial of the MDE Thesis can account for. In earlier work, we discovered that while people do not judge JoJo to be wholly blameworthy, they do not judge him to be blameless, either. And the data we present here seems to confirm this for other cases of moral deprivation. This suggests that while the MDE Thesis cannot be accepted as it stands, it has captured an important aspect of blameworthiness judgments. As for the Symmetry Thesis, we argue that it can be interpreted in a way that renders it consistent with our data.

In both cases—especially if we are to vindicate our interpretation of the Symmetry Thesis—it is an important question why people judge as they do. In our earlier work, we hypothesized that people judge JoJo as they do because of the difficulty JoJo would have in overcoming his upbringing and doing the right thing. After considering some alternatives, we argue that the very same hypothesis can help vindicate our interpretation of the Symmetry Thesis. This, we take it, both buttresses the plausibility of our claims about difficulty and represents an independent advance in our understanding of the roots of praise and blame.

Blinkered Badness

In the previous work mentioned, we explicitly explored folk intuitions on the JoJo case. In her original presentation and discussion of the case,
Susan Wolf offers a strong version of what we have labeled the MDE Thesis, claiming that people’s pretheoretic intuitions are that JoJo, because of his terribly deprived upbringing, is not morally responsible at all for his actions, despite their flowing from his deep self, i.e. his “true” or “real” evaluating or authenticating character. Her diagnosis for this alleged reaction is that JoJo’s deep self is normatively insane, i.e. he lacks the ability to recognize goodness, badness, and the difference between them. She then introduces a sanity—normative competence—condition into her theory of moral responsibility to account for such cases.

Insofar as Wolf is claiming a result on behalf of our pretheoretic intuitions, we decided to test her claim. Our findings were that, while people did judge JoJo to be less blameworthy than a control (JoJo’s presumably sane but nasty father), they still found him to be seriously blameworthy, assigning him an average blameworthiness score of 5 (where 7 was completely blameworthy and 1 was not at all blameworthy). Insofar as blameworthiness rides piggyback on responsibility (at least prima facie), JoJo is definitely viewed (pretheoretically) as morally responsible to some degree, contrary to Wolf’s strong version of the MDE Thesis. We took this at the very least to indicate that her adoption of a sanity condition was unmotivated.

But—and this is the interesting bit—JoJo was viewed as less blameworthy than he would have been without such a deprived background, so deprivation of this sort does seem to excuse to some extent. But what precisely was the relation of this deprivation to reduced blameworthiness? And could it be overcome?

Our initial diagnosis of JoJo’s reduced blameworthiness derived from his moral ignorance, not his normative insanity. It was not that he lacked the capacity to recognize right and wrong; it was, rather, that he had as a child been deprived of exposure to relevant moral alternatives. We attempted to explore these issues via the presentation of a third scenario in which JoJo was eventually exposed to moral alternatives but rejected them in favor of continuing to adhere to Daddy’s value system. Here,

---

6 On pp. 379–80, Wolf says “In light of JoJo’s heritage and upbringing . . . it is dubious at best that he should be regarded as responsible for what he does.” Later on p. 380, she says, “Our judgment that JoJo is not a responsible agent is one that we can make only from the inside” (emphasis ours).

7 Wolf, pp. 379–85.
surprisingly, there was no statistically significant difference between people’s reactions to the first and second JoJos. This might suggest, then, that it was their unfortunate formative circumstances that (slightly) mitigated their blameworthiness, and not their mere moral ignorance. But we remained unconvinced, arguing instead that the JoJos displayed a more fundamental and insidious type of moral ignorance than is usually discussed—an ignorance that expressions of ill will are wrong, not an ignorance of what specific act-tokens count as expressions of ill will—and that this sort of ignorance may not be displaced by mere exposure to moral alternatives (which would do nothing to counteract the thought that it is generally morally permissible to express ill will). What reduces the degree of their blameworthiness, we suggested, was not any sort of incapacity, though; rather, it was the difficulty of overcoming their childhood moral deprivations as adults, a difficulty that nevertheless assumes the possibility of success and so grants a basic normative capacity to them (a necessary condition to their being judged to be seriously blameworthy in the first place).

In our current round of experiments, we aimed to do two things. First, we wanted to see if we could duplicate the JoJo-type results in cases in which normative capacities were not at issue, and neither were any thoroughgoing childhood deprivations regarding the general wrongness of expressions of ill will. This would enable us to focus solely on the work being done by ignorance with respect to a specific sort of moral value. Second, we wanted to explore whether structurally identical positive cases would yield analogous results. We will explain this latter point in the next section of the essay, focusing here solely on the former.

New subjects were presented, at random, with one of the following two scenarios:

A. Tom is a white male who was raised in New Orleans. Growing up, he was taught to respect all people equally. Nevertheless, as an adult, he

---

8 We put this in terms of “ill will” in the original paper, but we recognize this to be a quite ambiguous notion. (For discussion, see David Shoemaker’s “Qualities of Will,” Social Philosophy and Policy 30 (2013): 95–120) We could have put this in less ambiguous terms as follows without losing the point. In standard cases of moral ignorance, the agent knows what properties make an act right or wrong but does not know whether some particular act-type instantiates those properties; JoJo’s ignorance, on the other hand, is about what properties make actions right or wrong in the first place, and so is much more profound.

9 \( n(A) = 84; n(B) = 84 \).
decided to become a proud racist, someone who believes that all non-white people are inferior and that he has a moral obligation to humiliate them when he gets a chance. At the age of 25, Tom moves to another town. Walking outside his home, he sees a black man who has tripped and fallen. In keeping with his moral beliefs, Tom spits on the man as he passes by.

B. Tom is a white male who was raised on an isolated island in the bayous of Louisiana. Growing up, he was taught to believe that all non-white people are inferior and that he has a moral obligation to humiliate them when he gets a chance. As an adult, he fully embraced what he’d been taught, becoming a proud racist. At the age of 25, Tom moves to another town. Walking outside his home, he sees a black man who has tripped and fallen. In keeping with his moral beliefs, Tom spits on the man as he passes by.

For each scenario, subjects were asked to rate Tom’s blameworthiness for spitting on the black man on a scale from 1 (“not all blameworthy”) to 7 (“completely blameworthy”).

Our prediction, in keeping with our results from the earlier paper, was that Tom$_B$’s upbringing and resultant ignorance with respect to the morality of racism would mitigate attributions of moral blameworthiness. The mean response to Tom$_A$ was 6.68, a very robust blameworthiness score. The mean response to Tom$_B$ was significantly lower at 5.4.$^{10}$ We conclude from this, again in keeping with earlier results, that certain sorts of moral blinders—i.e. childhood deprivations of exposure to moral truth—reduce blameworthiness assessments somewhat, but nowhere near completely.$^{11}$ And, once again, the fact that subjects viewed Tom$_B$ as seriously blameworthy reveals that they likely believed him not to lack the relevant normative capacities, i.e., sanity was not an issue. Finally, our

$^{10}$ Results were subjected to an independent samples t-test: $t(166) = 6.54, p < 0.001$ (two-tailed), $SD$ (ignorant) 1.53, $SD$ (non-ignorant) 0.92, Cohen’s $d = 1.01$.

$^{11}$ It might be thought that the vignettes do not do enough to establish sufficiently robust moral ignorance in Tom$_B$’s case, as anyone growing up in the modern era will surely have been exposed to alternatives through TV, radio, the Internet, or travel. Given that this might be an assumption of the subjects who read the vignettes, then it might have grounded their being pretty punitive. (Thanks to an anonymous reviewer for pressing this point.) This is a fair point, and it is worth controlling for. Nevertheless, we doubt it is doing much work for subjects, given that we stressed JoJo’s isolation and lack of exposure to alternatives in the original study, and we got nearly identical response levels to our JoJo doppelganger in the latest one.
speculative explanation remains quite plausible. We propose that the reason people judge Tom to be seriously blameworthy, despite the blameworthiness-reducing fact of his deprived childhood, is that, while it is more difficult for him to identify and do the right thing because of that childhood, it is nevertheless not overly demanding to expect him to do so.

Some new features we included buttressed this hypothesis by diminishing the plausibility of alternative explanations. First, we had Tom in both scenarios move to another town at the age of 25 in order to avoid the kind of ongoing isolated “preciousness” of the original JoJo case, and also to establish that he is an adult (with a fully-grown brain) who is making genuine moral decisions. His moving to another town would presumably also have given him opportunities to be exposed to moral alternatives. The question, then, was how that exposure would interact with what he had been taught as a child in people’s assessments of him. As it turns out, if he was taught the moral truth and rejected that as an adult, he was viewed as particularly blameworthy for spitting on the black man; he was viewed as less so if he had been taught the racist moral lie but simply did not reject it.

Second, we deliberately left open why it is that Tom either rejects or embraces what he has been taught. The reason was to allow for the possibility of any of the wide variety of sources of evaluations that occur in everyday life, and so not to privilege certain sources over others. Sometimes people evaluate after having pored over the reasons on both sides and determining which ones weigh more. Other times people evaluate after having seen a movie on the subject and without further deliberation. Other times people simply follow their intuitive hunches. Presumably all of these are methods that will preserve responsibility in assessors’ eyes.

Of course, the extent to which our proposal is plausible depends not only on the paucity of plausible alternatives, but on the inherent plausibility of the proposal itself. As has become clear to us since first introducing the proposal, more details about the nature of the difficulty in question are required before plausibility can be adequately assessed. However, we set this matter to the side momentarily in order to introduce the remainder of our data, which are relevant to our explication of the nature of the difficulty in question.

---

12 Thanks to two anonymous referees and to Shaun Nichols for pressing us to say more about the idea of difficulty in this context.
Thus far, we have been considering the effects of moral deprivation on assignments of blameworthiness. Especially given our interest in the Symmetry Thesis, the obvious question now is whether similar results obtain in cases of praiseworthy action. It is to that question that we now turn.

**Blinkered Goodness**

The Symmetry Thesis avers that blameworthiness and praiseworthiness are structurally analogous. If so, then it seems we should expect that childhood moral deprivations will reduce praiseworthiness in people’s eyes just as they do blameworthiness, at least for actions within the zone germane to the relevant deprivations. The basic thought here is rather compelling. It would seem that degrees of blame- or praiseworthiness both ought to track degrees of childhood-based moral ignorance in the same way: the less you know, the less you are “on the hook” in either case. If the Symmetry Thesis were true, childhood moral deprivations presumably ought to reduce moral responsibility all the way round.

The data, however, appear to undermine the Symmetry Thesis, at least when understood in this way. Further subjects were randomly assigned to one of the following two cases:\(^{13}\)

C. Tom is a white male who was raised in New Orleans. Growing up, he was taught to respect all people equally. Nevertheless, as an adult, he decided to become a proud racist, someone who believes that all non-white people are inferior and that he has a moral obligation to humiliate them when he gets a chance. At the age of 25, Tom moves to another town. Walking outside his home, he sees a black man trip and fall. Usually, Tom would spit on the man. But this time, Tom goes against his current moral beliefs, and helps the man up instead.

D. Tom is a white male who was raised on an isolated island in the bayous of Louisiana. Growing up, he was taught to believe that all non-white people are inferior and that he has a moral obligation to humiliate them when he gets a chance. As an adult, he decided to become a proud racist, embracing what he was taught. At the age of 25, Tom moves to another town. Walking outside his home, he sees a black man trip and fall.

\(^{13}\) n(C) = 85; n(D) = 83.
fall. Usually, Tom would spit on the man. But this time, Tom goes against his current moral beliefs, and helps the man up instead.

This time, subjects were asked to rate Tom’s level of praiseworthiness on a scale from 1 (“not at all praiseworthy”) to 7 (“completely praiseworthy”). The mean response to TomC was 4.28. The mean response to TomD was 5.40.14

There are several surprising conclusions one might draw here. First, TomC, who was raised with moral awareness but adopted racism as an adult, is only viewed as “somewhat praiseworthy” for going against those moral beliefs in doing the right thing. TomA—TomC’s twin who adhered to his adult-formed moral beliefs in spitting on the black man—was viewed as nearly completely blameworthy for doing so. Perhaps, then, the resistance to seeing TomC as more praiseworthy is due to his still being a racist.

This cannot explain all the data, however, because TomD is also a racist, and yet he is viewed as significantly more praiseworthy than TomC in going against his moral beliefs. It seems the only thing that could ground the difference in people’s assessments here is the difference in upbringings. But then here is the second surprising conclusion. As discussed above, it seems plausible that TomD’s moral ignorance would be generally mitigating: that if he did not know that his racism was wrong, then he could not know that his going against it was right. But apparently knowledge of the rightness of one’s actions is not viewed as necessary for praiseworthiness; indeed (and this is the truly surprising point) moral ignorance seems to be viewed as a virtue. Not only is the MDE Thesis being denied here, it is being turned on its head.

And thus we come to the third surprising feature. While moral ignorance reduces blameworthiness, it seemingly increases praiseworthiness. Importantly, this way of putting it suggests that the Symmetry Thesis as understood thus far is false, or at least highly problematic.

Resilient Symmetry

Nevertheless, we would like to explore the possibility of a different interpretation of the Symmetry Thesis that could allow for it to be maintained.

14 As before, results were subjected to an independent samples t-test: $t(166) = 4.18, p < 0.001$ (two-tailed), $SD$ (ignorant) 1.69, $SD$ (non-ignorant) 1.76, Cohen’s $d = 0.65$. 
in light of our results. To this point, we have understood the thesis to imply that factors reducing blameworthiness also reduce praiseworthiness, that negative and positive cases are symmetrical with respect to whether the relevant praise- or blameworthiness is reduced or increased (in comparison to some paradigm control case). It looks as if the Tom cases undermine this symmetry: the ignorance reducing Tom_B’s blameworthiness score actually increases Tom_D’s praiseworthiness score. Understood in this way, the relevant comparison point is the baseline at which Tom is not at all praise- or blameworthy. The right question to ask is: Did the feature in question move the degree of ______-worthiness away from or towards that baseline? (See Figure 1.1.)

Of course, one could simply accept the asymmetry. One might, for instance, attempt to explain it by drawing an analogy to the famous “Knobe effect.” This effect (named after Joshua Knobe) occurs when subjects are asked whether the side-effects of some agent’s actions were intentional.\textsuperscript{15} Subjects tend to say “yes” when the side-effects include something harmful, but tend to say “no” when the side-effects include something helpful. This appears to reveal an asymmetry in factual judgments of intentional action. Knobe’s own interpretation is that subjects’ judgments depend on their antecedent assessments of the normative status of the action’s side-effects, so that “bad” effects render the action intentional in subjects’ eyes, whereas “good” effects do not. Applied to our Toms, then, it may appear that a kind of reverse Knobe effect is revealed: moral ignorance via childhood deprivation seems to reduce attributions of responsibility when the actions are bad, whereas it increases such attributions when the actions are good. This could mean that factual judgments of responsible action also are dependent on antecedent assessments of the normative status of the action.

Given wide acceptance of the Symmetry Thesis, however, it seems that such a move would be overly hasty if it is possible instead to interpret the thesis in a way that is consistent with our data.\textsuperscript{16} Indeed, such an


\textsuperscript{16} While widely presupposed, the Symmetry Thesis is not accepted universally. For leading examples of asymmetricians, see Susan Wolf, Freedom Within Reason (Oxford: Oxford University Press, 1990); and Dana Nelkin, Making Sense of Freedom and Responsibility (Oxford: Oxford University Press, 2011). We have also already mentioned Doris and
interpretation exists: we propose, in contrast to the view represented by Figure 1.1, that negative and positive cases are symmetrically structured with respect to the direction in which various mitigating features shift the ______-worthiness judgments in relation to their controls. Understood in this way, the relevant comparison point would be the two endpoints of a continuous scale (see Figure 1.2).

So, with respect to the two endpoints (“completely blameworthy” and “completely praiseworthy”), the direction of movement from Tom\textsubscript{A} to Tom\textsubscript{B}, and from Tom\textsubscript{C} to Tom\textsubscript{D}, is symmetrical. What their sort of moral ignorance does, on this understanding of the relation, is move one more in the overall direction of complete praiseworthiness and away from complete blameworthiness.

If we understand the Symmetry Thesis in this way, then it might be preserved in light of our results. To do so, though, we must understand

Knobe’s arguments against a version of the Symmetry Thesis. Importantly, though, all of these theorists take their burden of proof seriously, and so admit the need to provide arguments against symmetry in blame and praise.
what the symmetry could consist in, i.e. what the explanation for the symmetrical movement would be. One possibility (continuing on the analogy with intentional action) stems from the work of Chandra Sripada, who has argued in favor of what he calls the “Deep Self Model” (DSM) of moral responsibility to provide a better explanation of the Knobe effect for action intentionality.\(^\text{17}\) What Sripada argues is that intentionality attaches to agents’ actions to the extent that their side-effects concord with the values and stable fundamental attitudes—the deep selves—that subjects attribute to them. If there is no concordance, goes the theory, subjects will be less likely to attribute intentionality to the agent’s actions. On this interpretation, then, subjects view the harmful side-effect in the Knobe studies as in concordance with the agent’s deep self and attribute intentionality to it thereby, but they view the helpful side-effect as in conflict with the agent’s deep self and so tend not to attribute intentionality to it thereby. This is how the DSM maintains a unified symmetrical explanation of the Knobe effect. Normative judgments of goodness and badness do not do the relevant work here; the different responses instead flow from perceptions of different structural underpinnings.

We might, then, appeal to the DSM to explain our results regarding responsibility. In particular, it will be useful for our purposes to explore whether the DSM alone can explain the results while also preserving the Symmetry Thesis, just as it purports to do with respect to the Knobe effect. One immediate problem with this is in figuring out just how the

concordance condition should be specified with respect to responsibility. Here is one plausible possibility: Agents will be viewed as more or less responsible for some action or attitude A solely to the extent that A is viewed as concording more or less with their deep selves.\footnote{Indeed, this would be the view Susan Wolf labels the “Deep Self View” of responsibility that she draws from Harry Frankfurt, Charles Taylor, and Gary Watson. See Wolf, pp. 373–9.} Of course, our scales were not put in terms of “responsibility”; rather, they were put in terms of praise- and blameworthiness. But perhaps we might articulate the DSM in terms of those labels as follows: agents will be viewed as more or less ______-worthy for A solely to the extent that A is viewed as concording more or less with their deep selves. Call this the DSM\textsubscript{BP}. On this application of the DSM, one will be viewed as more blameworthy in correspondence with the perception that A more closely concords with one’s deep self (and A is bad), whereas one will be viewed as more praiseworthy in correspondence with the perception that A more closely concords with one’s deep self (and A is good). Can this explain the difference between our various Toms?

At issue is why moral ignorance via childhood deprivation moves one in a positive direction a full step up from control cases in the eyes of subjects. The DSM\textsubscript{BP} would have us believe that this is because recognition of that sort of moral ignorance affects subjects’ attribution of the relevant actions to the agents’ deep selves. But this does not seem to fit the results of our study. According to the DSM\textsubscript{BP}, given that Tom\textsubscript{B} is less blameworthy than Tom\textsubscript{A}, it must be that Tom\textsubscript{B}’s action concords less with his deep self—similarly for Tom\textsubscript{D} vs. Tom\textsubscript{C}. But neither view seems motivated, precisely because in each pairing the Toms appear to have identical deep selves. For example, both Tom\textsubscript{A} and Tom\textsubscript{B} fully embrace their racism as adults, and both act on it upon seeing the fallen black man. The only difference is in their upbringings. So why should we (or any subject) therefore think Tom\textsubscript{B}’s action is less reflective of his deep self than Tom\textsubscript{A}’s?

One response here would be to suggest that perhaps Tom\textsubscript{B} lacks, or at least is deficient in having, a deep self, given the limited range of moral alternatives to which he was exposed in his upbringing. This might then explain why subjects judge him to be less blameworthy: his action is less concordant with his deep self than Tom\textsubscript{A}’s action insofar as Tom\textsubscript{B} does
not have as robust a deep self as Tom\textsubscript{A}. Unfortunately, making this move undermines the DSM\textsubscript{BP} explanation of the positive cases, for if a deprived moral upbringing renders one’s deep self less robust or more deficient, and so mitigates attributions of responsibility, then it ought to render Tom\textsubscript{D} less praiseworthy than Tom\textsubscript{C} in the eyes of subjects. But this is the opposite of our results.

Perhaps we can get a more plausible version of the DSM\textsubscript{BP} if we start on the praiseworthy side of the map. How might we explain the differing assessments of Tom\textsubscript{C} and Tom\textsubscript{D} on the DSM\textsubscript{BP}, given that both Toms were stipulated as identically embracing their racism and so would seem to have identical deep selves? Perhaps the thought is this: Tom\textsubscript{D}’s upbringing somehow made him less committed to the moral beliefs with which he was raised, such that when he goes against them he is more likely to be expressing his actual deep self than is Tom\textsubscript{C}. In other words, given their differences in upbringing, Tom\textsubscript{C}’s action is viewed as more “out of character,” more anomalous, than Tom\textsubscript{D}’s. This thought could then be extended to the negative cases as follows: Tom\textsubscript{B} is less blameworthy than Tom\textsubscript{A} because his upbringing somehow renders his racist action more anomalous than Tom\textsubscript{A}’s. Tom\textsubscript{B} is not “really” committed to his racism, subjects might think, at least in the way that Tom\textsubscript{A} is, so Tom\textsubscript{B} is less blameworthy thereby.

This interpretation of the DSM\textsubscript{BP} preserves the Symmetry Thesis, but how plausible is it? It is unclear to us why subjects would consider Tom\textsubscript{D}’s commitment to his moral beliefs to be any less serious than Tom\textsubscript{C}’s (or that Tom\textsubscript{B}’s is less serious than Tom\textsubscript{A}’s). Indeed, might not their kind of restricted ideological upbringing make them more committed to the belief system into which they had been indoctrinated? Might not this indoctrination create, if anything, a deep self more in line with the belief system than an upbringing without it? To the extent that these core “moral” principles were instilled in a way that bypassed the ignorant Toms’ rational, evaluative stance, they are likely to be resistant to such evaluations, much in the way religious belief with its source in childhood indoctrination is often difficult to expunge.

To be clear, we do not reject the possibility of reading a deep self view into the results here. It could well be that subjects really are viewing Tom\textsubscript{D}’s commitments as less attributable to his deep self than Tom\textsubscript{C}’s. Nevertheless, given the vignettes as stated, we have no evidence for this approach. Certainly, we welcome any future attempts to see whether
differential deep self commitments are doing any work here. Until then, though, we believe we are licensed in assuming either that the DSM_{BP} supports an asymmetrical approach to blame- and praiseworthiness or that its symmetrical approach is unmotivated and (at least initially) implausible. Focusing on attributions to the deep self alone does not look as if it will help us.

In our discussion of the MDE Thesis, we proposed that people’s reactions to JoJo and Tom_{B} might stem from an appreciation of the difficulty those agents would have in doing the right thing. We believe that this idea can helpfully be extended to likewise explain our results regarding the Symmetry Thesis:

**Difficulty Hypothesis**: Moral ignorance resulting from childhood deprivation functions symmetrically in both negative and positive cases (moving assessments up the single scale of blameworthiness to praiseworthiness in relation to the control) in virtue of the difficulty agents are viewed as having in overcoming their morally deprived upbringing to grasp the relevant moral reasons.

On this hypothesis, as before, Tom_{B} is viewed as less blameworthy than Tom_{A} in light of how difficult it would be for him to go for a moral alternative not included in his morally blinkered upbringing. Our further suggestion is that Tom_{D} is viewed as more praiseworthy than Tom_{C} in light of how difficult it in fact was for him to go for a moral alternative not included in his morally blinkered upbringing. This way of viewing the matter easily explains the results while avoiding the dual implausibility of thinking that subjects think (a) there is a difference in the deep selves for either of the pairings, and (b) childhood indoctrination actually renders one’s deep self more open or oriented toward moral alternatives than non-indoctrinated childhoods. If this is right, then the Symmetry Thesis may not after all be undermined by cases of moral ignorance in upbringing.

As noted earlier, however, we need to be clear about just what the Difficulty Hypothesis amounts to. In particular, we need to say something more about the nature of the difficulty in question. For one thing, one might worry that talk of difficulty just collapses into talk of capacity; perhaps, after all, Wolf was right to appeal to capacities. For another, it might be thought that we are suggesting that doing something inherently difficult (like, say, lifting something heavy) is, in itself, sufficient for praise. We agree that this would be implausible.

First, even if we grant that talk of difficulty is another way of talking about capacity, at the very least our results suggest that capacity talk must
be scalar. After all, both Tom_B and Tom_D are still viewed as ______-worthy for what they have done to a significant extent. This means that even though such ______-worthiness is thought to be affected by moral deprivation and ignorance, it is so only to some degree, so that if the Toms are viewed as being incapacitated in some respect it would only be an incapacity by degree. This does not sound, however, like traditional talk of capacity, which is typically taken as either obtaining or not; and it is not clear what the details of this take on capacity would consist in. At the very least, to reintroduce talk of capacity would require some difficult explicatory work about its re-envisioned nature that we are unable to assess sight unseen.

Regardless of whether our claims can somehow be adapted to a kind of capacity-talk, our interest is not in capacities themselves, but in the difficulty of exercising certain capacities, where we measure this against a baseline of what comparable agents might be expected to do. Though there is some historical precedence for including difficulty in exercising a capacity in the criteria for moral responsibility, the focus has typically been on volitional capacities.\textsuperscript{19} It is difficult for the unwilling addict, for example, to resist taking the drug because his counter-desire, his craving for the drug, is so strong. To the extent we cut him some slack, then, we may do so because we think it was just too difficult for him to overcome that volitional obstacle, where the same would be expected of other similarly situated agents. (Of course, this way of putting it might raise capacity talk once more, for perhaps the unwilling addict really is incapacitated with respect to volitional obstacles of that strength, or perhaps he partially lacks a “meta-capacity” to exercise his volitional capacities.\textsuperscript{20})

The difficulty we appeal to is analogous to that just mentioned, but differs in that it concerns perceptual, rather than volitional capacities.\textsuperscript{21} Indeed, there is no reason to think that our Toms are struggling against their own desires and inclinations otherwise to do what is right. Rather, it is just harder


\textsuperscript{20} One might read Harry Frankfurt in the former way. See his “Freedom of the Will and the Concept of a Person,” in Harry Frankfurt, The Importance of What We Care About (Cambridge: Cambridge University Press, 1988), pp. 11–25.

\textsuperscript{21} By “perceptual” capacity, we mean whatever capacity allows apprehension of moral qualities in specific cases. We are not assuming that moral knowledge is fundamentally perceptual or empirical.
for our Toms to “see” what the right thing to do is, given their morally depraved upbringings. When they do, it is surprising, for we tend to think that most people from their background would not have seen the light.

As an analogy, suppose that I have been shown the famous image of the “duckrabbit” repeatedly since childhood (see Figure 1.3), and I have been taught over and over that what I am looking at is a duck, and only a duck. When, as an adult, I meet you, and you insist that the image can also be seen as a rabbit, it will be no surprise if, given my upbringing, I have a very hard time coming to see it as a rabbit. Certain features of the picture have been drilled into me as exclusively salient (e.g., the bill), so it is quite difficult for me to come to see other features (e.g., the little “rabbit mouth” indentation on the back of the duck’s head) as salient in my assessment of the image. Suppose, however, that someone else present, who was raised just as I was to see the duck, is able to see the rabbit quite quickly once told about it. Especially if we have taken my inability to see the rabbit as representative of the baseline, we are likely to be impressed with this person. We will probably applaud his ability to see past what he is used to.

Of course, again, one might suggest that this is best understood in terms of capacities—perhaps he has a stronger “meta-capacity” to exercise his perceptual capacities than I do—but this is not necessary. Regardless of where we come down on capacity-talk, the point remains that this person’s ability to see the rabbit is impressive, and speaks highly of him as a perceiver. Our suggestion is that, on analogy, we are more likely to praise TomD because we are impressed with his ability to “see” past what he is used to, morally speaking. This way of thinking about difficulty also responds to the worry that praiseworthiness might attach to any old difficulty. We are merely advancing the Difficulty Hypothesis to explain cases of moral ignorance given childhood deprivation, and

![Figure 1.3 Duckrabbit.](image-url)
what it appeals to is specifically perceptual difficulty in grasping moral reasons. We take no stand on whether it applies to volitional difficulty, nor other sorts of difficulty for that matter.

Revisiting Huck

Our primary aim in this essay has been to bring some empirical results to bear on the MDE and Symmetry theses. While the MDE Thesis looks false in light of our results, there is a weaker, scalar, version of it that may be defensible. And while it may seem as if our results cause real worries for the Symmetry Thesis, it too may be defended. What inclines us toward the latter stance in both cases is the unified defense that may be provided them by the Difficulty Hypothesis.

We conclude where we began, with Huckleberry Finn. Huck has been raised to believe that stealing someone’s property is immoral, and that slaves are people’s property. As he grows he continues to mouth and even embrace the racist judgments instilled in him by his family and community since childhood. But when he is finally presented with an opportunity to turn Jim in, he balks, going against his moral beliefs and thus opting for hell instead. He is represented by our TomD, and people view him not only as praiseworthy, but as more praiseworthy than someone just like him but without the childhood moral deprivation.

There has been a lot of philosophical ink spilled (or at least a lot of words processed) on the Huck Finn case in recent years. Our aim has not been to resolve the many issues raised in these discussions; rather, it has been the far more modest one of providing some much-needed empirical background to them. Far too often, discussion of the case proceeds by stating what our pretheoretic intuitions about Huck are. Ours is the only study we know of that attempts to determine just what those intuitions consist in. People do view him (or someone very much like him) as quite praiseworthy, although not completely so.

Further, we have attempted to show just how much of a role the moral
deprivations of someone-like-Huck’s upbringing contribute to people’s
assessments of his praiseworthiness. It indeed matters that he has been
raised the way he has, moving him up the scale of praiseworthiness a full
step over someone who does what he does without such a background.
But one surprising feature of the study is that his increased praiseworthi-
ness seems disanalogous to how people view our JoJo lookalike, who is
less blameworthy—less responsible?—for his bad deeds than a morally
undeprived doppelganger.

Indeed, here is the precise rub for determining whether the results of
our study negatively impact or ultimately reinforce the Symmetry Thesis:
childhood moral deprivations are often thought to mitigate responsibil-
ity itself. Being more or less blameworthy may simply be viewed as
translating to being more or less responsible. If responsibility is indeed
scalar in this fashion, then our results suggest that there is in fact an
asymmetry between negative and positive cases: morally deprived
upbringings are viewed as decreasing one’s responsibility in the negative
realm and increasing one’s responsibility in the positive realm. This
would truly be a surprising result.

Less surprising, but perhaps no less interesting, would be a different
interpretation. Responsibility may not be scalar in the way just suggested;
instead, perhaps it is either not scalar at all (one either is responsible for
something or one is not), or it is scalar, just not in line with the scalar
nature of blame- and praiseworthiness (perhaps it is scalar in line with
various capacities one may have or exercise by degrees, where this does
not affect degrees of blame- or praiseworthiness). If this is the case, then
the Symmetry Thesis might be reinforced. If we think of blame- and
praiseworthiness on a single scale, then (where full-blown responsibility
would attach to one’s actions at any point on the scale, say) the moral
deprivations of childhood could be viewed symmetrically in negative and
positive cases as moving one away (by roughly the same amount) from
the completely blameworthy endpoint.

While we have not taken a definitive stand either way here, our
previous discussion on the JoJo case does provide some explanatory
ammunition if one adopts the second approach. What could well be
moving the assessments of ______-worthiness up the scale away from
complete blameworthiness is that people view childhood moral depriva-
tions as making moral perception more difficult for their agents, so that
adhering to, or overcoming, the judgments ingrained by those deprivations will predictably yield assessments at a different degree than their non-deprived counterparts. Just as we cut JoJo some slack for doing what it would have been difficult for him not to see as wrong, we also admire Huck for having done what it was difficult for him to see as right. Or at least we admire Huck in a different way, or to a different degree, than we would those who more easily saw what he did not because of their more privileged upbringing. As his internal monologue suggests, it is a serious and genuine struggle for him to reject “morality” in the way he does. That he did so and got it right (despite what he still thought) is worthy of real praise, apparently. It is a surprise.

There is, as always, much more work to be done on these issues. In particular, it would be valuable to know more about the reasons why our positive Toms went against their moral beliefs. Again, we deliberately left this part open so as not to beg the question against any particular theory of relevant moral reasons, but in doing so we also left it open that the Toms might have changed their minds by accident, or under the influence of various non-rational forces. (We doubt this is how the cases were interpreted, but it is certainly possible.) It would also be interesting to know whether it is the mere isolation of their upbringing or the specific moral facts from which they have been deprived that is doing the work on people’s different intuitions. To test this, one might have TomC and TomD raised in different degrees of isolation but now indoctrinated with the same set of moral facts, that people are to be treated equally, etc.; and then have them both act in accordance with their moral beliefs in helping the fallen man up. Whether or not there is a difference in people’s moral intuitions, we would learn something interesting. In addition, the degree of praiseworthiness people would assign may be compared to the degrees assigned in our present cases, so that we may know more about how people assess right-doing in line with moral upbringing. Indeed, finding out how people assess these cases will likely tell us even more about how we ought to view the Symmetry Thesis, and so could ultimately generate real philosophical payoff.23

23 We are very grateful to the folks at Yale’s Experiment Month for carrying out our proposed study on these issues. We are also grateful to Don Callen for serving as advisor to the study, and to Tamler Sommers and Michael McKenna for discussion of some of the ideas herein.
The Cognitive Mechanisms of Intolerance

Do our Meta-Ethical Commitments Matter?

Jennifer Cole Wright, College of Charleston
Cullen B. McWhite, College of Charleston
Piper T. Grandjean, University of Wyoming

Now more than ever, peaceful coexistence requires open, compassionate dialogue between cultures that have very different sets of beliefs, values, and practices. Unfortunately, such dialogue can encounter intolerance that threatens not only to shut it down but also to generate strong negative reactions—even violence—towards divergent cultures (Haidt et al., 2003; Skitka et al., 2005; Skitka and Mullen, 2002; Wright et al., 2008).

Recent scholarship has identified potential sources for this intolerance. Particularly highlighted is the role that people’s moral beliefs and values, especially those strongly held, play in generating negative interpersonal reactions. People report being significantly less supportive of moral diversity than other forms of diversity (Haidt et al., 2003; Wright et al., 2008), especially when encountered in contexts where intimate and/or frequent interaction is likely (e.g., dates, room-mates, workplace). And

Correspondence concerning this chapter should be sent to: Jennifer Cole Wright, Department of Psychology, College of Charleston, 57 Coming Street, Charleston, SC 29424, email: wrightjj1@cofc.edu.
even though people are more tolerant of moral diversity in cultures other than their own, this is true only when they believe it to result from different factual—not moral—beliefs (Wainryb, 1993; Wainryb et al., 1998; Wainryb et al., 2004).

Believing a disagreement to involve moral beliefs and values is a powerful predictor of intolerance in people of all ages, resulting in less tolerant attitudes towards divergent beliefs and values and less willingness to interact with, help, share resources with, or sit close to those who have them (Wright, 2012; Wright et al., 2008). Similarly, people’s moral conviction predicts their intolerance for divergent beliefs and values, their unwillingness to seek resolution, and their suspicion for legal/political processes perceived as supporting those divergent beliefs and values (Mullen and Skitka, 2006a, 2006b; Skitka et al., 2005; Skitka and Mullen, 2002).

Meta-Ethical Commitments

What is it about people’s moral beliefs and values that drive this lack of tolerance? One plausible explanation comes to us from meta-ethics. Some philosophers hold that people are moral objectivists (Blackburn, 1984; Brink, 1989; Mackie, 1977; Smith, 1994), taking the position 1) that moral judgments are truth-apt, and 2) that their truth is mind-independent (and thus not determined by an individual’s or group’s beliefs, values, customs, etc.). Such a position creates the space for moral beliefs and values to be either (objectively) right or wrong. Given that morality is a normative domain, providing guidance on how things ought to be, this problematizes tolerance: to the extent that people are inclined to think the “right” moral beliefs and values are their own, they will view any that differ as wrong, misguided, even dangerous and immoral (e.g., Rachels and Rachels, 2009; Wong, 1984, 2006; cf. Snare, 1992). Thus, if people are (as many assume) moral objectivists, they are likely to display intolerance of divergent moral beliefs and values.

But are people moral objectivists? Though people’s meta-ethical commitments have been a topic of much debate in philosophical circles (Blackburn, 1984; Brink, 1989; Harman, 1975; Mackie, 1977; Rachels and Rachels, 2009; Shafer-Landau, 2003; Smith, 1994; Wong, 1984, 2006), it has received less empirical attention. Indeed, it was not until recently that researchers have begun to systematically explore people’s
metaethical commitments (Goodwin and Darley, 2008, 2010, 2012; Sarkissian et al., 2011; Krettenauer, 2004; Wright et al., 2012; Wright and Sarkissian, 2013; though see also Fishkin, 1984).

Of course, given the research highlighting the role morality plays in generating intolerance, it seems reasonable to assume that people are objectivists. And there is plenty of (indirect) evidence for this. People of all ages distinguish between moral and non-moral issues, viewing moral issues as more universally relevant, transcending cultural particulars, and applying more globally to human valuesbehavior (Nucci and Turiel, 2000; Killen and Nucci, 1995; Nucci, 1981; Smetana, 1981, 1983; Wainryb et al., 1998, 2004; Turiel, 1983, 1998).

Moreover, moral considerations are treated as more important than other considerations (Kohlberg, 1969, 1986; Piaget, 1932; Rest, 1979). Moral transgressions are viewed as wrong even in the absence of rules or the presence of social sanction (Smetana, 1981, 1983; Stoddart and Turiel, 1985; Turiel, 1983) and are seen as more serious, less response-dependent, and more severely punishable than social transgressions (Turiel, 1983, 1998; also Davidson et al., 1983; Goodwin and Darley, 2008; Nucci, 1981; Smetana, 1981, 1983; Smetana and Braeges, 1990)—findings that hold cross-culturally (Nucci et al., 2000; Nucci et al., 1983; Song et al., 1987; Turiel et al., 1988).

In summary, there is a body of findings that support the hypothesis that people are moral objectivists, viewing divergent moral beliefs and values as violations of objective truths. Yet there is other research that brings this into question. Nichols (2004) found that while most of the people questioned about moral transgressions gave objectivist responses, some did not. Even more interestingly, Goodwin and Darley (2008, 2010, 2012) found that while people gave more objectivist responses for moral issues than other issues, they were nonetheless internally pluralistic in their metaethical stance, giving objectivist responses for some moral issues but not others.

One potential explanation for this pluralism is that people gave nonobjective groundings for some “moral” issues (identified as such by the researchers) because they did not view them as moral issues. Nonetheless, Wright et al. (2012) found a similar pattern of pluralism when they allowed participants to self-identify moral issues. They also found strong evidence that this pluralism was genuine—i.e., not merely the result of conceptual confusion or different views about morality.
Given this surprising meta-ethical variability, what conclusions should we draw about the relationship between people’s meta-ethical commitments and intolerance? Since previous research has failed to adequately disentangle domain classification (moral/non-moral) and grounding (objective/non-objective), we can only hypothesize that people’s meta-ethical commitments will be strongly related to their tolerance for divergent beliefs and values, along with their willingness to interact with or help the people who endorse them; that, while people will express greater intolerance for divergent moral than non-moral beliefs and values generally, this intolerance will be greatest for objectively-grounded moral issues. Support for this hypothesis was found by Goodwin and Darley (2012)—asking participants to rate their discomfort for a room-mate with divergent moral beliefs and values revealed a significant correlation \( r = .90 \) between grounding and discomfort.

But this leaves an important question yet unanswered: why would people display meta-ethical pluralism (given that it seems an incoherent position)? Given the proposed relationship between grounding and tolerance, our hypothesis is that meta-ethical pluralism serves an important psycho-social function: namely, modulating the level of permissible choice and dialogue about moral issues, both within and between socio-cultural groups. Viewing a moral issue as objectively grounded removes it from the realm of legitimate personal/social negotiation (i.e., individual and/or social attempts to condone it will be deemed unacceptable, and censorship/prohibition will be supported). Viewing a moral issue as non-objectively grounded, on the other hand, allows people to acknowledge its moral significance (i.e., that it is not simply a personal matter), while at the same time maintaining room for choice, dialogue, and debate—thus, social censorship/prohibition will be viewed less favorably. The studies reported below are the first that we know of to test this hypothesis.

**Study 1**

**PARTICIPANTS**

Seventy-two undergraduate students who were enrolled in Introduction to Psychological Science courses participated in this study for research credit. The data from nine participants were eliminated due to incomplete surveys. Of the remaining sixty-three participants, 78% reported being
female; 86% Caucasian, 2% African-American, 3% Asian-American, 6% Hispanic.

MATERIALS

Participants were given a survey with forty-five issue-statements (e.g. “People should not cheat on their spouse,” “People should wash their bodies regularly,” see Table 2.1), seventeen of which came from Goodwin and Darley (2008) and Wright et al. (2012) and the remainder from Wright et al. (2008). Statement presentation was randomized to control for order effects.

Domain classification. For each issue-statement, participants were asked to categorize the issue, choosing the best fit from the following: “personal choice/preference,” “social convention/norm,” “moral,” or “scientific fact.” Participants were not given further instructions about these categories; not only does previous research (Wright, 2012; Wright et al., 2012) support the contention that people have a competent conceptual grasp of these categories, but it would have been difficult to provide definitions that would not have influenced participants’ classification.

Self-classification was chosen over researcher classification (the method typically employed) because research on the relationship between moral beliefs and values and tolerance (e.g., Wright, 2010; Wright et al., 2008, 2012) has uncovered significant disagreement, between and within age-groups, about what issues belong in the moral domain. This does not appear to be due to conceptual confusion or different views about morality—qualitative analysis of people’s explanations for their classifications suggests they share largely the same ideas about what morality is and what characteristic features it has (at least at its “core”; cf. Haidt, 2012). Rather, disagreement typically involved whether, for a given issue, moral features were present and/or should be prioritized over other features. Therefore, when assessing people’s meta-ethical commitments, it seemed better to not assume the moral status of any issue a priori.

---

1 Some issues might fit into multiple categories. Therefore, we asked participants to choose whichever category was the best fit, capturing the features that were most salient and/or took priority.