UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
The Relationship Between Social Cognition and Moral Reasoning

Permalink
https://escholarship.org/uc/item/1qd6j747

Author
Uttich, Kevin Thomas

Publication Date
2012

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

The Relationship Between Social Cognition and Moral Reasoning

By

Kevin Thomas Uttich

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Tania Lombrozo, Chair
Professor John Kihlstrom
Professor Elliot Turiel

Fall 2012

The Relationship Between Social Cognition and Moral Reasoning

Copyright 2012

By

Kevin Thomas Uttich

Abstract

The Relationship Between Social Cognition and Moral Reasoning
By
Kevin Thomas Uttich
Doctor of Philosophy in Psychology
University of California, Berkeley
Professor Tania Lombrozo, Chair

What is the best way to think about the relationship between Social Cognition and Moral Reasoning? Past psychological research has treated the relationship between the two as one-way with descriptive information from social-cognitive capacities impacting normative moral judgments. However, some resent findings have challenged this account. Across three sets of studies this dissertation examines these challenges, and asks whether the relationship between Social Cognition and Moral Reasoning is best understood as bi-directional. Further, two sub questions are investigated; would a bi-directional relationship require a drastic revision of our understanding of social cognition and does the influence of norm information depend on the type of norm involved?

The first set of studies examines the influence of norms on Theory of Mind judgments. Theory of Mind, the capacity to understand and ascribe mental states, has traditionally been conceptualized as analogous to a scientific theory. However, recent work in philosophy and psychology has documented a "side-effect effect" suggesting that moral evaluations influence mental state ascriptions, and in particular whether a behavior is described as having been performed 'intentionally.' This evidence challenges the idea that theory of mind is analogous to scientific psychology in serving the function of predicting and explaining, rather than evaluating, behavior. In three experiments, we demonstrate that moral evaluations do inform ascriptions of intentional action, but that this relationship arises because behavior that conforms to norms (moral or otherwise) is less informative about underlying mental states than is behavior that violates norms. This analysis preserves the traditional understanding of theory of mind as a tool for predicting and explaining behavior, but also suggests the importance of normative considerations in social cognition.

The second set of studies looks at the use of norms in explanation. In explaining behavior, people often refer to mental states such as beliefs or desires. But in some cases, behavior can also be explained by direct appeal to norms, moral or otherwise (e.g., "she returned the wallet because it was the right thing to do"). We investigate whether and when norm explanations are accepted (Experiment 1) and produced (Experiment 2) to better understand the relationship between norms and mental states in explaining behavior. In particular, do norm explanations assume particular mediating mental states, like the agent's knowledge of the norm? We find that participants frequently accept and produce norm explanations for behavior when the behavior matches the norm, even when the agent's belief about the norm is incorrect. The findings contribute to a growing body of work suggesting that mental state inferences and reasoning are not detailed and automatic, but instead remain relatively underspecified for the purposes of many everyday judgments.

Finally, the third set of studies investigates whether people accept "ethical explanations," explanations that cite moral norms (and not merely people's *beliefs* about moral norms) to account for social-historical changes, such as the abolition of slavery or the extension of voting rights to women. An ethical explanation for women's suffrage, for example, might cite the injustice of withholding the right to vote on the basis of sex. Such explanations pose a challenge to dominant accounts of explanation, which propose that explanations cite causes or descriptive generalizations. In two experiments, we find evidence that people do accept and provide ethical explanations, and that variation in ratings and production of these explanations is correlated with two separate types of meta-ethical commitments: belief in moral objectivism and in moral progress. These results suggest that some people accept ethical explanations because their particular meta-ethical commitments lead them to conceptualize moral norms in a way that allows them to serve as legitimate explanations. The findings also shed light on variation in moral beliefs across individuals despite reasonable consistency in moral beliefs within individuals.

Taken together the results of these studies argue that the relationship between social cognition and moral understanding is indeed bi-directional. In addition to the influence the information from social cognitive judgments can have on our moral reasoning, our normative understanding can provide a source of information useful in making social-cognitive judgments. The results also indicate that a drastic revision of our understanding of social cognition is not necessary and while all norm types seem to have an influence, different norm types have unique relationships.

Table of Contents

Title Page
Copyright
Abstract 1
Table of Contents i
Acknowledgements ii

Chapter 1: Introduction **1**

Chapter 2: Norms inform mental state ascriptions: A rational explanation for the side-effect effect **6**
Introduction *6*
Experiment 1 *9*
Experiment 2 *12*
Experiment 3 *17*
General Discussion *20*
Tables *29*

Chapter 3: Norms as explanations for behavior. **30**
Introduction *30*
Experiment 1 *32*
Experiment 2 *37*
General Discussion *41*
Figures *46*
Tables *50*

Chapter 4: Ethical Explanations, Moral Objectivity, and Moral Progress **54**
Introduction *54*
Experiment 1 *58*
Experiment 2 *66*
General Discussion *72*
Figures *77*
Tables *79*

Chapter 5: Conclusion **83**

Chapter 6: References **88**

Acknowledgments

To begin, I would like to thank my advisor and dissertation chair Tania Lombrozo for her support, her patience, and her willingness to venture with me down this line of research. It was only through Tania's guidance that my half formed ideas turned into research projects that I am truly proud of. I could not have asked for a better mentor. I would also like to thank my collaborator George Tsai. Having the opportunity to work with and learn from people like you is what makes research so enjoyable. I would like to thank my committee members Elliot Turiel and John Kihlstrom. I was incredibly fortunate to get your feedback on my work, and also learn so much about moral judgment and social cognition through my graduate career from such distinguished experts. Also, thanks to the McDonnell Foundation and NSF grant BCS-0819231 to Tania Lombrozo for supporting this research.

Next, I would like to thank my lab mates Nicholas Gwynne, Michael Pacer, Joseph Williams, Joe Asterweil, Elizabeth Seiver, Caren Walker, Chris Lucas, Daphna Buchsbaum, Chris Holdgraf, Casey Sullivan, Christine Fawcett, and Christine Schafer Willis for their input and insights. Thanks to the Moral Psychology group at Berkeley, particularly Audun Dahl, Dylan Murray, Eugene Chislenko, and Olga Antonenko, as well as members of the Moral Development seminar, for making it easy and enjoyable to keep up with all the new research and learn the many things I should have already known. Thanks to my wonderful research assistants who helped me collect, organize and analyze data Kathie Pham, Megan Wynn, Jesse van Fleet, and Govind Shantharam. I would also like to thank all the members of the Concept and Cognition lab who helped me complete these projects particularly, Ania Jaroszewicz, Paul Rogerson, Emily Wei, Jared Lorince, David Oh, Lisa Qin, Christina Botros.

I'd like to thank all those who took the time to provide helpful feedback on drafts or presentations of the studies included in this dissertation particularly Joshua Knobe, Edouard Machery, Lori Markson, Jennifer Cole Wright, Bertram Malle, Steve Guglielmo, Joanna Korman, Andrew Monroe, Kyle Dillon, Jonathan Phillips, Fiery Cushman and Steven Sloman. I would also like to thank everyone at the University of Chicago who helped me find my way to psychology research, particularly Amanda Woodward, Jennifer Sootsman-Buresh, Jessica Heineman-Pieper, Bertram Cohler, Susan Goldin-Meadow, Susan Levine, Kristi Schonwald and everyone at the University of Chicago Language Development Project.

Thanks to all of my friends who helped me along the way particularly Zach Rosner, Auden Dahl, Eric Walle, Alex Main, Wilbur Putney Williams III, Ellen Wallace, James Morgante, Greg Pierce, Robert Leake, Jeremy Fowler and James Leahy.

Lastly I would like to thank my parents who taught me to love learning and always supported my interests, and my sisters Katie and Erin for keeping me on my toes. I'd like to thank my dogs, Louise and Witz. Having extra reason to wake up early in the morning, and to take a long, mind clearing, walk in the evening was crucial to my productivity and mental well-being. I owe so much to my amazing wife Laura who supported and encouraged me every step of the way. There is no way I could have made it through graduate school without her. Finally I'd like to thank my favorite person in the whole world, my amazing daughter Elizabeth, who gave me extra motivation to finish writing this damn thing in a timely manner.

Chapter 1: Introduction

Psychological research concerning social cognition and moral understanding has examined these two areas independently and treated them as distinct and fundamental psychological capacities. What is the best way to think about the relationship between these two capacities? Accounts of social-cognitive capacities, including attribution theory (Malle, 2004), mental state ascription (Gopnik, 1999; Wellman, 1992) and causal reasoning (Kelley, 1967; Woodward, 2004) have tended to consider the relationship as one-way, with descriptive information from our social-cognitive capacities impacting the normative moral judgments we make (Malle and Nelson, 2003). However, recent work (Knobe, 2003a, 2010) has suggested that this typical understanding may miss aspects of the relationship. In this dissertation I investigate whether the relationship between social cognition and moral understanding is better described as bidirectional, with information from moral reasoning impacting social-cognitive judgments as a typical part of the process. This possibility for bi-directionality comes as a result of taking seriously an argument put forth by Searle (2001) that claims prescriptive norms provide reasons for agents to act in accordance. Because prescriptive norms provide a reason for action, they provide both a reason to predict conforming behavior and a reason to infer other motivations, for example mental states, when the norm is violated. In this dissertation I investigate what this argument about prescriptive norms would mean for the social-cognitive inferences people make and the explanations people provide.

Social Cognition and Moral Understanding as Distinct Capacities

Social-cognitive capacities as used here are broadly defined as capacities people use to make sense of others and themselves (Fiske and Taylor, 2007). While social cognition is a wide field with many different theoretical varieties, this dissertation will focus on an aspect common to many of the views, namely the explanation and predication of the behavior of others. As used here, the term "social-cognitive capacities" will refer to the capacities that serve a role in explaining or predicting the behavior of others. Primarily this will refer to people's capacity to infer mental states and their capacity to draw on these mental states and other factors to explain behavior. These two parts of social cognition are intertwined: mental state information can be used to explain and predict behavior, and behavior can in turn be used to infer mental states. In particular, the social-cognitive capacities that will be discussed include attributing mental states such as beliefs, intentions, or desires and explaining behavior.

Most accounts of social-cognitive capacities have not included any role for moral understanding (Knobe, 2010). Some of the classic work on attribution (Heider, 1958; Jones and Davis, 1965; Kelley, 1967; Skowronski and Carlston, 1987) emphasized the importance of atypical (and hence counternormative) behavior in guiding judgment. However, the norms in question in this research were descriptive, concerning what typically happens or is statistically likely, rather than prescriptive, indicating what should or ought to happen.

The capacity for moral understanding as it is being used here refers to our capacity to reason or make judgments about what one ought to do or what should occur. This capacity is normative in that there is an ideal or value attached to the judgments and the role of this capacity is evaluative. While there is some disagreement in what the scope

of the moral domain includes, whether it concerns only issues of harm, fairness and justice (Turiel, 1983) or whether the scope widens to include issues of loyalty, authority, and sanctity (Haidt and Joseph, 2004), there is general agreement that the moral domain concerns at least evaluations of good or bad on harm, justice and fairness issues. Moral understanding as used here includes moral reasoning, moral judgments and moral evaluations, or the process, decisions and beliefs about whether an individual or an action is morally good or bad.

Concerning the other direction of the relationship, research has well established that the information provided by social-cognitive capacities impacts moral reasoning. Countless studies have demonstrated the impact of intentional and causal attributions on moral judgments (Cushman and Young, 2011; Darley and Shultz, 1990; Fincham & Jaspers, 1979; Kohlberg, 1969; Malle & Nelson, 2003; Piaget, 1965/1932; Shultz, Wright, & Schleifer, 1986; Weiner, 1995). These studies indicate that in addition to the resulting outcome and the specific behavior that is being evaluated, moral judgments also take into account mental state information, like whether an action was intentional, or what the person believed or desired. For example, if a gun is shot and a person is killed, in evaluating the situation and making a moral judgment people will take into account social-cognitive information like whether the trigger was pulled intentionally or accidently bumped, or whether the shooter desired to shoot the person or wanted to hit a target, or whether the shooter knew there was a person in the path of their shot or thought that the area where they aimed was clear. However, social-cognitive reasoning and moral reasoning are distinct and separate capacities under these accounts. Under these types of accounts, when information from moral reasoning impacts social-cognitive judgments it means that the social-cognitive judgments have been corrupted or biased by the moral information, and the process was incorrect or incomplete.

Challenges to Distinct Capacities Accounts

However, there are reasons to think that this typical account might be incorrect or perhaps incomplete, and the best way to describe the relationship between social cognition and moral reasoning is actually bidirectional. In particular, this dissertation will examine two examples where normative information seems to impact social-cognitive capacities in their normal course of functioning. First, there are results of studies where different judgments of intention are made when a side effect is good versus when a side effect is bad. Second, prescriptive norm information is often seen in explanations of behavior, like "because it is the right thing to do" as an explanation for why a lost wallet was returned.

Norms and inference

Recently philosopher Joshua Knobe has presented some evidence suggesting moral considerations can impact social-cognitive judgments, specifically intentional action judgments. This phenomenon is known as the side-effect effect or the Knobe effect. In his initial studies Knobe presented cases where a chairman either helped or harmed the environment as a foreseen side-effect of making profits. When participants were asked if the chairman intentionally harmed the environment, 82% said yes. However, when the new program's side effect was to *help* the environment, only 23% of participants said the chairman intentionally helped the environment (Knobe, 2003a). Since the harm and help vignettes seem to differ only in the moral valence of the side effect, Knobe interpreted these results as suggesting that *moral* considerations somehow

influence intentional action judgments, and perhaps the broader category of Theory of Mind judgments, judgments concerning the mental states of others. These results have been replicated with different methodologies (Knobe 2003a, 2004; Knobe and Mendlow 2004; Machery, 2008), across cultures (Knobe and Burra A., 2006), and with preschool children (Leslie, A., Knobe, J. & Cohen, A., 2006).

Knobe argues that these results are evidence that Theory of Mind competencies are shaped by the role Theory of Mind judgments play in evaluating behavior, be it in assessing moral responsibility or assigning praise and blame, suggesting ‘‘ . . .moral considerations are actually playing a role in the fundamental competencies underlying our use of the concept of intentional action” (Knobe, 2006). This interpretation suggests that the influence of moral norms on social-cognitive capacities necessitates a complete revision of both the model and the primary function of social-cognition.

Others interpreted the effect as the result of a bias in Theory of Mind judgments caused by the norm information. On this view, moral evaluations are not contained within Theory of Mind judgments, but can exert an extraneous influence. For example, conversational pragmatics (Adams and Steadman, 2004a), the desire to blame an agent for a negative outcome (Mele, 2001; Malle & Nelson 2003), or an emotional reaction (Nadelhoffer, 2004b) could lead participants to (mistakenly) describe the side-effect as having been brought about intentionally. This view does not require a drastic revision of the function of Theory of Mind or other social-cognitive capacitates, but requires the claim that moral evaluations don’t really interact with social-cognition but have an external, biasing effect.

These findings illustrate a way in which moral norms impact social cognitive judgments concerning mental states. However, these findings also raise many questions about the exact nature of this interaction and what the interaction means about our model of social cognition.

Norms and explanation

Developmental research also provides some examples of prescriptive norm information influencing judgments children make about an agent’s behavior (Kalish, 1998). For example, young children predict that an agent will conform to a norm, even if the norm is unknown to the agent or conflicts with the agent’s own preferences. However, older children and adults predict that when norms and preferences conflict, preferences will most often win out (Kalish & Cornelius, 2007; Kalish & Shiverick, 2004). While this and other adult work suggests we develop an ability to balance belief and desire information with norms, there is also research suggesting that even in adults reasoning about beliefs might not be so simple. Specifically Apperly et al. (2006) found that participants are slower to infer belief when they are not explicitly asked to track it, suggesting belief inferences are not automatic. There is also work suggesting belief reasoning is effortful, where participants with more working memory constraints are less effective in applying Theory of Mind (Lin, Keysar & Epley, 2010).

Most philosophical and psychological theories of explanation have difficulty accounting for the impact of prescriptive norms in explanations. Explanations like “She returned the wallet because it was the right thing to do” intuitively seem like acceptable explanations, but it is unclear the exact way they are explanatory. While accounts of explanation vary, most suggest that explanations cite one or more causes (e.g., Salmon 1984; Woodward, 2003), provide underlying mechanisms (e.g., Bechtel, 2008; Craver,

2007; Darden, 2006; Ahn and Kalish, 2000), and/or appeal to descriptive generalizations, such as physical laws (see Woodward, 2010 for review from philosophy; see Lombrozo, 2011 for review from psychology). It's not immediately apparent how prescriptive norms could fulfill any of these definitions. Many philosophers, historians, and social scientists have doubted whether explanations that cite a moral or ethical quality can actually be adequate explanations of social and historical facts (Leiter, 2001; Harman, 1977; Williams, 1985), but some have argued that ethical explanations can legitimately account for social and historical change (Cohen, 1997). One theory of explanation where an explicit role for norm information has been carved out is found in the work of Bertram Malle. In his work Malle distinguishes an alternative type of explanation that instead cites the factors that bring about reasons, which he refers to as "causal history of reasons" explanations (1999). Malle cites "That's the cultural norm" as one potential type of causal history explanation. The use of causal history explanations varies with the situation being explained. Of particular importance for our purposes are the results of a study by O'Laughlin and Malle (2002), which found that causal history explanations were provided more often for group behavior than for individual behavior.

In this dissertation I investigate whether the relationship between social cognition and moral understanding is better described as bidirectional by taking seriously the idea that prescriptive norms provide reasons for agents to act in accordance with the norm (Searle, 2001), and investigate what this would mean for the social-cognitive inferences people make and the explanations people provide. If prescriptive norms provide reasons for agents to act in a particular way, then the presence of a prescriptive norm should provide a source for explaining and predicting future behavior as well as a reason to infer a relatively strong cause that is personal to an agent, like a mental state, when an agent violates a prescriptive norm.

In addition to this overarching issue, I will address two sub-questions. First, if there is an influence of normative considerations on social-cognitive judgments, does this require a drastic revision of our understanding of social cognition or can we account for this normative influence while largely preserving our existing models of social cognition? Additionally, I will address whether the influence of norm information differs based on the type of norm. More specifically, I will examine whether moral norms have a unique influence, or whether the influence is similar to what is observed with descriptive and conventional norms.

Dissertation Overview

In the following three chapters I will examine how our understanding of norms about what ought to, should, or typically does happen impacts our social cognitive reasoning. Chapter two will present evidence that we use prescriptive norm information to infer mental state information used in intentional action judgments. Chapter three will present evidence that norms sometimes serve as explanations for an agent's behavior, independent of an agent's belief. Finally, in chapter four I will demonstrate that ethical or moral norms can serve as explanations for social-historical events, and present evidence that their use is related to people's beliefs about the nature of morality.

To foreshadow the results, we find support for a bi-directional relationship driven by norms. In chapter two we find that while there is an impact of moral status on judgments of intentional action, the impact is not exclusive to moral norms and the effect also impacts predictions of future behavior. In chapter three we show that norms are

suitable explanations for behavior and that their use as explanations is not just shorthand for a causal process that includes belief, but that not all norms had the exact same effect. Finally in chapter four we find that explanations that cite a moral claim are a part of people's reasoning about social-historical change, and that the use of these explanations is related to people's belief about the nature of the moral world.

Chapter 2: Norms inform mental state ascriptions: A rational explanation for the side-effect effect[1]

Chapter two examines a line of research that finds that moral considerations influence judgments of intentionality (Knobe, 2006), and suggests the relationship between social cognition and moral understanding is better described as bidirectional. In this chapter we investigate this phenomenon, look at whether the effect on intention is really best described as a moral influence, and examine whether a change in the directionality of the relationship between moral understanding and a social-cognitive capacity would require a drastic revision of our understanding of the function and purpose of the capacity.

Introduction

Consider sitting at a commencement address and thinking, "that speaker must love to wear billowy black gowns." This attribution is odd, because we know that academic norms dictate commencement attire. But upon viewing someone dressed in full regalia at a café, it might be appropriate to infer an underlying mental state, such as a false belief that it is commencement or a desire to look scholarly, because in this situation the academic norm does not apply. These examples illustrate that norms inform mental state ascriptions. More precisely, prescriptive norms provide reasons for acting in accordance with those norms (Searle, 2001), with the consequence that norm-conforming behavior is relatively uninformative about underlying mental states: one needn't observe norm-conforming behavior to infer underlying reasons to obey the norm. In contrast, norm-violating behavior *is* informative about underlying mental states, as there must be a reason behind the norm-violating behavior, and moreover the reason must be sufficiently strong to outweigh the reason(s) to observe the norm.

The capacity to understand and attribute mental states is often characterized as a theory of mind (e.g. Gopnik, 1999; Wellman, 1992). Like a scientific theory, Theory of Mind (ToM) posits unobserved entities (internal states) to support explanation and prediction. Knowing that a man in a café desires to appear scholarly, for example, can explain eccentric attire, and supports predictions about whether he is more likely to smoke a pipe or a cigar. But for the commencement speaker, eccentric attire is better explained by appeal to a conventional norm, and smoking habits are better predicted from baserates. These observations suggest that norms *should* inform mental state ascriptions if reasoners are to be effective "intuitive scientists" (Kelley,1967), and if ToM is to accomplish the functions of predicting and explaining behavior.

This paper explores the relationship between norms and mental state ascriptions by considering the relationship between prescriptive norms – both moral and conventional – and ascriptions of intentional action. Previous work suggests that ascriptions of intention have an impact on moral evaluations (e.g. Malle & Nelson, 2003). For example, an intentional killing is typically judged a murder, while an unintentional killing is considered manslaughter (e.g. California Penal Code). But recent findings suggest that the reverse may likewise hold – that moral evaluations can influence ascriptions of intentional action (Knobe, 2003a, 2006). Specifically, Joshua Knobe has

[1] The material in chapter 2 was co-authored with Dr. Tania Lombrozo.

uncovered an intriguing asymmetry in judgments concerning whether actions that brought about morally good versus bad side effects were performed 'intentionally', a phenomenon known as the side-effect effect or the Knobe effect. Consider the following vignette, which Knobe presented to participants in his initial studies:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'
>
> The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was harmed.

When participants were asked if the chairman intentionally harmed the environment, 82% said yes. However, when the new program's side effect was to *help* the environment, only 23% of participants said the chairman intentionally helped the environment (Knobe, 2003a). Because the chairman expressed indifference to the side effect in both vignettes, judging either side effect intentional violates previous accounts of intentional action, which identify intent and desire, along with skill and foresight, as prerequisites to intentional action (Malle and Knobe, 1997). Moreover, the harm and help vignettes seem to differ only in the moral valence of the side effect, which suggests that *moral* considerations somehow influence ToM judgments.

The side-effect effect has been replicated with different methodologies (Knobe 2003a, 2004; Knobe and Mendlow 2004; Machery, 2008), across cultures (Knobe and Burra A., 2006), and with preschool children (Leslie, A., Knobe, J. & Cohen, A., 2006). While a variety of explanations for the effect have been offered (Knobe, 2006; Machery, 2008; Adams and Steadman, 2004a; Nadelhoffer, 2004b), no single proposal successfully accounts for all the data collected to date (Pettit and Knobe, 2009).

Broadly speaking, responses to the side-effect effect have fallen into two distinct camps, which we call the 'Intuitive Moralist' view and the 'Biased Scientist' view. The Intuitive Moralist view takes the effect as evidence that ToM competencies are shaped by the role ToM judgments play in evaluating behavior, be it in assessing moral responsibility or assigning praise and blame. For example, Knobe writes that "…moral considerations are actually playing a role in the fundamental competencies underlying our use of the concept of intentional action" (Knobe, 2006). This interpretation not only challenges the idea that the influence of ToM judgments on moral judgments is one-way, but also the idea that the function of ToM is to predict and explain behavior – instead, ToM may be a multi-purpose tool partially shaped by its role in moral evaluation.

The Biased Scientist view instead suggests that the effect results from a bias in ToM judgments. On this view, moral evaluations are not contained within ToM judgments, but do exert an extraneous influence. For example, conversational pragmatics (Adams and Steadman, 2004a), the desire to blame an agent for a negative outcome (Mele, 2001; Malle & Nelson 2003), or an emotional reaction (Nadelhoffer, 2004b) could lead participants to (mistakenly) describe the side-effect as having been brought about

intentionally. Here ToM capacities are still regarded as the product of an 'intuitive scientist', but the particulars of the Knobe scenarios lead to results the intuitive scientist cannot accept. Judgments are consequently altered to generate a more acceptable result. This view preserves the traditional function of theory of mind, adding the claim that moral evaluations can have a biasing effect.

We propose a third way of explaining the side-effect effect and of understanding the relationship between ToM and moral judgment. Perhaps moral judgments inform ToM judgments, but not because moral considerations partially constitute or bias ToM concepts. Rather, as suggested in the introduction, actions that violate norms (e.g. harming the environment) provide a basis for ascribing counter-normative mental states and traits to an agent, whereas actions that conform to norms do not. This asymmetry in ascribed mental states and traits is sufficient to in turn generate the asymmetric judgments that characterize the side-effect effect.

We call our proposal the 'Rational Scientist' view to emphasize that inferring mental content on the basis of a behavior's relationship to norms (moral or otherwise) makes sense if the goal of ToM is to support prediction and intervention. We suggest that people can make use of information about the agent being evaluated, situational factors, applicable norms, and so on to draw initial or 'baseline' mental state and trait inferences (call them 'MST1'). After observing the agent's behavior, mental state and trait ascriptions can be updated, yielding MST2. Whether or not a behavior is considered intentional is a function of MST2. While norm-conformance provides little evidence to change MST2 from MST1, norm-violating behavior suggests mental states or traits strong enough to outweigh reasons to obey the norm, and as a result MST2 will be quite different from MST1. When the CEO knowingly proceeds with a plan that will harm the environment, for example, MST2 may supply the desire or intention component required by the Knobe and Malle (1997) model of intentional action (for related arguments about differences in desire across conditions see Gugiellemo and Malle, 2009; and Sripada, 2009 for the relationship between disposition and self).

The Rational Scientist view differs from the Intuitive Moralist view in preserving the traditional function of theory of mind: prediction and explanation. Our approach concedes that moral judgments influence ToM, but this influence is seen as *evidential*, not *constitutive*. In other words, moral norms affect ToM ascriptions by influencing mental state ascriptions, but such ascriptions are not inherently evaluative. The Rational Scientist view also differs from Biased Scientist views in regarding the influence of moral judgment on ToM as a rational strategy for achieving the function of ToM, and not as a bias or extraneous pressure. While our view differs from many contemporary explanations for the side-effect effect, it shares important elements with classic ideas in attribution, such as the Correspondent Inference Theory of trait attribution (Jones and Davis, 1965), the cue-diagnosticity approach to trait attribution (Skowronski and Carlston, 1987), and the Covariation ANOVA Model (Kelley, 1967), many of which emphasize the importance of atypical (and hence counternormative) behavior in guiding judgment (see also Gugiellemo and Malle, 2008, Sripada, 2009, Sripada and Konrath, 2010, Holton, 2010).

In this paper we test the Rational Scientist view as a hypothesis about the relationship between moral evaluation and theory of mind. First, we examine whether the asymmetric ascriptions of intentional action in previous demonstrations of the side-effect

effect stem from the side effects' *norm status* or their *moral status*. In previous cases, "harm" scenarios involved bad side effects that resulted from norm-violating actions, while "help" scenarios involved good side effects that resulted from norm-conforming actions. Experiments 1 and 2 deconfound moral status and norm status to examine what drives the side-effect effect: norm status, as predicted by the Rational Scientist view, or moral status, as predicted by the Intuitive Moralist and Biased Scientist views. Experiment 1 additionally examines whether effects of norm status are restricted to moral norms or extend to conventional norms. Second, we examine whether norm-violating actions are indeed more informative than norm-conforming actions when it comes to positing mental states and traits that support prediction. This is the focus of Experiment 3.

To preview our results, we find that the asymmetry in the side-effect effect results from the side effects' norm status, that the side-effect effect extends to conventional norms, and that norm-violating behavior supports stronger predictions about future behavior than norm-conforming behavior. These findings offer strong support for the Rational Scientist view, and provide a way to understand the relationship between ToM and moral judgment.

Experiment 1

In focusing on norm status and mental state inferences, rather than on moral evaluations, the Rational Scientist view makes a few unique predictions. First, because the Rational Scientist view argues that what drives the side-effect effect is the relationship between norms and behavior, not the moral status of behavior or outcomes itself, the Rational Scientist view predicts that judgments of intentional action should vary when the norms in a situation vary, even if a behavior and its outcome remain the same. Second, because the Rational Scientist view argues that the asymmetry in the side-effect effect results from mental state inferences licensed by norm-violations, the Rational Scientist view predicts that the effect should extend to non-moral norms, such as conventional norms. While other studies have provided evidence that the side-effect effect is not limited to moral cases (Machery, 2008), they have not focused on conventional norms or on asymmetries arising from norm-conformance versus norm-violation.

Experiment 1 investigates both predictions using vignettes in which an agent acts to bring about an intended, main effect with a foreseen side effect. While the agent's action and the side effect are held constant across conditions, norm status is varied by introducing industry standards. For example, one set of vignettes involves a CEO who pursues an action with a 25% chance of causing environmental harm, but where the industry standard for pursuing a plan with environmental risk specifies that the probability of harm must be either 45% or less (making the behavior *norm-conforming*) or 5% or less (making the behavior *norm-violating*).[2] While norm status varies across

[2] Each scenario involves multiple behaviors, some of which could potentially be considered norm-violating in the norm-conforming condition (e.g. stating a lack of care for the environment) or norm-conforming in the norm-violating condition (e.g. pursuing a plan that will increase profits). Because the behavior the scenarios were designed to test particular norms (e.g. one should not harm the environment without sufficient reason), we continue to refer to the scenarios as simply "norm-conforming" or "norm-violating"

conditions, the probability of harm (25%) is held constant, and the environmental harm always occurs. In matched vignettes involving a conventional norm, the CEOs actions will change the color of a manufactured product to black, where the color change is either norm-conforming (the product is conventionally darker than blue) or norm-violating (the product is conventionally lighter than blue). If the Rational Scientist view is correct, participants should judge it more appropriate to say a side effect was brought about intentionally in the *norm-violating* cases than in the *norm-conforming* cases, for both moral and conventional norms, no matter that matched cases involve identical side effects.

Participants

Participants were 300 University of California-Berkeley undergraduates who received the questionnaire as part of a larger packet completed for course credit. There were 12 conditions with 25 participants in each condition.

Materials and Procedure

Participants read vignettes involving an agent who acted to bring about a main effect with a foreseen side effect. The side effect could be norm-conforming or norm-violating, and the relevant norm could be *moral* or *conventional*. To match vignettes as closely as possible, the side effect was held constant across pairs of scenarios, with norm status manipulated by stipulating an industry standard in each vignette. Specifically, the moral vignettes always involved a side effect with a 25% probability of occurring, but the stipulated norm was to proceed when the side effect had less than a 45% chance of occurring, making the action norm-conforming, or when the side effect has less than a 5% chance of occurring, making the action norm-violating. However, the side effect always in fact occurred. Below are sample vignettes:

> [Moral norm Condition] A regulatory agency for the Gizmo industry exists in order to provide environmental standards even though it does not have the authority to ensure compliance with these standards. This regulatory agency has established an environmental standard, which states that a company may only start new programs if the chance of environmental harm due to the program is under 5% [45%].
>
> The vice-president of a company in the Gizmo industry went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but there is a 25% chance that it will also harm the environment. The industry standard is to only start programs of this type when the chance for harm is under 5% [45%]."
>
> The chairman of the board answered, "I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program."
>
> They started the program. As it happened, the environment was harmed.
>
> [Conventional norm condition] The convention in the Gizmo industry is for Gizmos to be a dark color. Specifically, the convention is to make Gizmos

---

depending on whether the agent's behavior violates the norm that varies across paired vignettes.

that are colored darker than blue.

The vice-president of a company in the Gizmo industry went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will result in our Gizmos being colored black. The convention is to make Gizmos colored darker than blue, so we would be complying with the convention." [The convention is to make Gizmos colored lighter than blue, so we would be violating the convention.]

The chairman of the board answered, "I don't care at all about the color of the Gizmos. I just want to make as much profit as I can. Let's start the new program."

They started the program. As it happened, the Gizmos were black, colored darker than blue.

Participants were then asked to rate how appropriate it would be to say that the side effect was brought about intentionally, providing ratings on a 1 to 7 scale, with 1 being "not at all appropriate," 7 "very appropriate," and 4 "neither appropriate not inappropriate." For the sample vignettes above, they were asked: "How appropriate is it to say the CEO intentionally harmed the environment [The chairman of the board intentionally made Gizmos colored darker than blue]?".

In addition to varying the nature of the norm (moral, conventional) and the side effect's norm status (conforming, violating), there were three distinct sets of vignettes, one involving a CEO and included above, one involving a doctor (DR) and one involving a trucking company (TRUCK). There were thus 12 distinct vignettes, with participants randomly assigned to a single vignette.

Results and Discussion

Participants' ratings of whether it is appropriate to say that the agent brought about the side effect "intentionally" (see Fig. 1.1) were analyzed using an ANOVA with three between-subjects factors: norm status (2: conforming, violating), norm type (2: moral, conventional), and vignette version (3: CEO, DR, TRUCK). This analysis revealed a main effect of norm status ($F(1,288)=12.828$, $p<.01$), with norm violating side-effects receiving higher ratings than norm conforming side-effects. There was also a main effect of vignette ($F(2,288)=11.705$, $p<.01$), with average ratings in the DR Vignette lower overall. There was no interaction between norm status and norm type ($F(1, 288)=2.269$, $p=.133$), suggesting the effect was comparable for both norm types. In all 12 conditions the average ratings for the norm-violating side effects were numerically higher than those for the norm-conforming side effect.[3]

These results suggest that in evaluating whether an outcome was brought about intentionally, participants consider the relationship between behavior and norms, and not

[3] Because the CEO vignette involving a moral norm has been the focus of so much debate, we ran a post-hoc t-test comparing intentional action ratings as a function of norm status for just this vignette, revealing a non-significant effect (4.8 versus 5.2, $t(48)=-.91$, $p = .37$). However, a replication restricted to this condition with 431 participants revealed that those in the *norm-violating* (5%) condition generated significantly higher ratings of intentional action (4.96, sd=1.66) than did those in the *norm-conforming* (45%) condition (4.42, sd=1.73; $t(429)=-3.28$, $p<.01$).

merely the behavior or its outcome. Thus the asymmetry observed in the side-effect effect does not depend specifically on a difference between "good" and "bad" actions or outcomes as most versions of the Intuitive Moralist and Biased Scientist views would predict, but rather on the difference between norm-conforming and norm-violating actions. Moreover, the importance of *norm* status as opposed to *moral* status is reinforced by the fact that the effect is also observed when the norms in question are conventional. Like moral norms, conventional norms provide reasons for action, establishing an asymmetry in the mental states one can infer (MST2) on the basis of norm-conformance versus norm-violation.

Experiment 2

While Experiment 1 is consistent with the Rational Scientist view and makes the case that the side-effect effect extends beyond moral norms, other explanations for the data are possible. In particular, an advocate for the Intuitive Moralist or Biased Scientist view could argue that stipulating a norm influences judgments of intentional action by establishing whether a side effect is good or bad, with participants' own evaluations of "goodness" or "badness" ultimately responsible for judgments, not norm status per se. This concern is plausible in light of the fact that the scenarios involved uncertain side effects about which participants had little prior knowledge. Providing norms may have effectively *taught* participants what counts as good and bad in the course of the experiment. While this concern already concedes a role to norms, Experiment 2 replicates the key findings with side effects for which participants have strong, antecedent moral judgments.

Experiment 2 thus employs vignettes with side effects that are likely to generate strong moral evaluations with or without experimental context, and includes an assessment of participants' own evaluations of the moral status of the side effects. To manipulate norm status while keeping the moral status of side effects constant, an agent's actions are embedded in a context with typical moral norms (the 'superhero' context) or a context with reversed norms (the 'supervillain' context). So, for example, the side effect of accelerating global warming should be norm-violating for a superhero and norm-conforming for a supervillain, but is likely to be judged morally bad by all participants.

Because the Rational Scientist view claims that norm status drives the side-effect effect by determining which mental states are ascribed to an agent, it predicts that changing a vignette's context (superhero versus supervillain), and therefore the norms with respect to which the agent operates, should influence judgments of intentional action. For example, a supervillain who *decelerates* global warming is violating a supervillain norm to cause harm, so one can infer that the supervillain must have had a reason to bring about this (good) outcome that was sufficiently strong to outweigh reasons to conform to supervillain norms. This (good) outcome should therefore support stronger ascriptions of intentional action than a (bad) outcome that conforms to supervillain norms, such as *accelerating* global warming. In contrast, because both alternative views focus on moral status and participants' moral evaluations of the side effects, they would presumably predict that responses will track participants' moral evaluations of the side effects, irrespective of vignette context. That is, an agent who accelerates global warming should be judged to have done so intentionally and one who

decelerates global warming should not, irrespective of whether the agents are superheroes or supervillains.

Participants

Participants were 96 University of California-Berkeley undergraduates who received the questionnaire as part of a larger packet completed for course credit. There were 8 participants in each of 12 conditions.

Materials and Procedure

Participants read a vignette about an agent who acted to bring about an intended main effect and a foreseen side effect, where the side effect was either morally good or morally bad. However, the agent was embedded either in a context with typical norms concerning morally good and bad action (the 'superhero' context) or in a context with reversed norms (the 'supervillain' context). Participants were asked to take the perspective of an assistant to a superhero or supervillain and to evaluate the actions of an agent who was being considered for a promotion. Below is an example of a vignette from the supervillain condition, involving a harmful side effect:

> There is a Supervillain that has a group of evil henchmen who work for him. The Supervillain and his henchman are the badest of the bad, never passing up a chance to spread malice and evil. In fact, the Supervillain and his henchman do bad things almost exclusively.
>
> You are the assistant to the Supervillain. Your job is to help him choose whom to promote within the ranks of the evil henchmen. The job of the evil henchmen is to do maximum evil at every opportunity. To make your decision, you've planted secret microphones and video cameras to observe the henchmen in action. Below is an episode that you've learned about concerning Bob, a henchman eager to move up the ranks whom you will be asked to evaluate:
>
> A rookie henchmen said to Bob: "Sir, we have a plan to rob a bank by releasing neurotoxins in the air, temporarily causing anyone in the vicinity of the bank to be incapacitated. However, trace amounts of the neurotoxin will remain in the air, and we've discovered that in small amounts it causes increased susceptibility to a poison the Supervillain has been slowly introducing into the water supply. Exposure to the neurotoxin will cause even more people in town to become ill from the water and the reactions will be more severe. It will greatly harm the town's quality of life."
> Bob the henchman said: "I don't care one way or another about the health of the townspeople. I just want to steal as much money as I can. Let's go ahead with the plan."
>
> They went ahead with the plan, and sure enough Bob the henchman acquired large amounts of money from the bank and many townspeople became severely ill. The health of the townspeople was thus harmed.
>
> Your boss, the Supervillain says: "I need to choose someone to promote up

> the villain ranks. I have a lot of henchman to choose from. All of them have proven that they are capable of doing bad, but I want someone who is truly evil. What about Bob? I heard he robbed a bank, acquired a large amount of money and made a bunch of townspeople ill.

The corresponding vignette for the superhero condition involved stopping a bank robbery by using the neurotoxins, and as a side effect harming the health of the townspeople. For both the superhero and supervillain conditions, in the corresponding "help" side-effect condition participants were told: "Exposure to the neurotoxin will cause fewer people in town to become ill from the water and the reactions will be less severe. It will greatly increase the town's quality of life." As in the harm cases, the agent claimed to not care about the foreseen helpful side effect, and the effect always occurred as expected. The other two sets of vignettes involved destroying the UN (supervillain's goal) or a Supervillain's compound (superhero's goal) using a giant mirror and thereby accelerating (harm) or decelerating (help) global warming, and blowing up a dam (supervillain's goal) or building a dam (superhero's goal) with negative (harm) or positive (help) consequences for the town's fishing industry.

Participants were asked to maintain the perspective of the assistant to the supervillain (superhero) and were told to "respond to the following questions as if the supervillain (superhero) asked you." They were then asked, "How appropriate is it to say Bob intentionally harmed the health of the townspeople?" and provided ratings on a 1 to 7 scale as in Experiment 1. Participants made additional judgments (see Table 1.1) to examine whether the context manipulation effectively altered judgments concerning the agent's behavior, and to examine the inferences participants drew about the agent on the basis of the context and side effect. To verify that participants' own norms corresponded to the superhero context, with the "help" side effect judged good and the "harm" side-effect judged bad, participants were asked to respond to additional questions "from your own personal perspective (as if you were telling a friend about Bob instead of responding to the supervillain as his assistant)" (see Table 1.1).

Participants were randomly assigned to one of 12 conditions, the result of crossing side-effect moral status (2: harmful, helpful), context (2: superhero, supervillain), and vignette version (3: bank robbery, global warming, fishing).

Results and Discussion

The critical dependent measure was participants' evaluation of whether it is appropriate to say that the agent brought about the side effect "intentionally." We analyzed ratings using an ANOVA with three between-subjects factors: side-effect valence (2: harmful, helpful), context (2: superhero, supervillain), and vignette version (3: bank robbery, global warming, fishing). This analysis revealed a main effect of side-effect valence ($F(1,83)=7.17$, $p<.01$), with harmful side-effects receiving higher ratings than helpful side-effects, as well as the predicted interaction between side-effect valence and context ($F(1, 83)=20.91$, $p<.01$; see Fig. 1.2). There were no other significant effects. In the superhero context, the results replicated past demonstrations of the side-effect effect, with the harmful side effect receiving higher ratings for intentional action than the helpful side effect. However, this pattern was not observed for the supervillain context; in fact, the ratings for the helpful side effect were numerically higher than those for the

harmful side effect. Judgments about whether the main effect was intended were uniformly high (5.84, s.d. = 1.52), and did not vary as a function of condition.

Findings involving the remaining dependent measures are summarized in Table 1.1, which indicates the means for each judgment as a function of SE valence and context, as well as significant main effects and interactions. First, consider the judgments made from the perspective of the assistant to the superhero or supervillain. The fact that participants rated heroes more likely than villains to bring about good effects in the future (a & b) confirms that participants understood the intended, typical behavior for agents in each community. More reassuring, the significant interaction between SE valence and context for judgments about the side effect, the agent, blame versus praise, and promotion (e, f, g, & h) all suggest that participants effectively adopted the intended perspective, and were able to evaluate the agent with respect to the stipulated norms.

The questions about the agent's future behavior in relation to an average candidate (c & d) were intended to test the hypothesis that norm-violating behavior is more informative than norm-conforming behavior in the sense that it provides evidence to alter predictions from baseline, which should correspond to the predictions for an average agent (4 on the 7-point scale). That is, MST2 should differ more from MST1 for norm-violation than for norm-conformance. This predicts that agents who conform to norms (a helping hero or a harming villain) should generate judgments very close to 4, while agents who violate norms (a harming hero or a helping villain) should differ from 4, with harming heroes more likely to harm and less likely to help in the future, and helping villains less likely to harm and more likely to help. While this pattern or results was obtained for the heroes, it was not for the villains. It may be that some participants assumed that a norm-violating agent would compensate for the norm-violation – for example, that a supervillain who helped would make up the help with future harm. Because these findings are difficult to interpret, Experiment 3 examines the influence of norm-violation and norm-conformance on future prediction more directly.

Finally, consider the judgments that were made from the perspective of the participant. Unsurprisingly, participants judged good side-effects good and bad side-effects bad; heroes good and villains bad; and praised heroes more than villains, with greater praise for bringing about good side effects. These findings reinforce that participants' *own* moral evaluations were consistent across conditions, and that differences in ascriptions of intentional action stemmed from the relationship between an agent's behavior and the norms with respect to which that behavior was evaluated, not the moral 'goodness' or 'badness' of the actions or outcomes themselves.

While these additional dependent measures serve principally to confirm background assumptions, they also provide an opportunity to examine the relationship between these judgments and ascriptions of intentional action. Ratings for whether the side effect was brought about intentionally correlated significantly with the valence of the side effect from the perspective of the vignette ($r=-.39$, $p<.001$), with higher ratings for intentional action corresponding to ratings that the side effect was more negative. However, an equivalent relationship was not observed for the participants own perspective ($r=-.17$, $p=.12$), again suggesting that participants' own moral judgments played little role in ascriptions of intentional action.

These results provide evidence for the Rational Scientist view over alternatives. While the superhero cases replicate previous findings, reversing the norms with a

supervillain context had a corresponding effect on ascriptions of intentional action. This reversal is predicted by the Rational Scientist view. While a participant's norms and an evaluated agent's norms may often be the same – especially if participants consult their own norms as a default – the two can diverge when there's evidence that an agent subscribes to different norms, as in our supervillain context. The norms attributed to the agent's in turn determine mental state ascriptions, because only norms that apply to an agent can supply that agent with a reason to act in accordance with the norm, and hence generate the evidential asymmetry that we suggest drives the side-effect effect. In contrast, because the Intuitive Moralist view, as well as most versions of the Biased Scientist view, suggest that participants are tracking moral status or are influenced by their own moral understanding, these views predict that ascriptions of intentional action should track a participants' own moral evaluations, not those of an arbitrarily stipulated context within which the evaluated agent is operating.

Additionally, Experiment 2 addresses a potential concern about Experiment 1: that judgments in Experiment 1 were only influenced by norms because participants did not have a prior basis for making an evaluation about the valence of the side-effect. In Experiment 2, participants had clear judgments about the status of the side effect, and these judgments were not influenced by context.

Experiments 1 and 2 thus make the case for the role of norm status rather than moral status in generating the side-effect effect. However, there are two potential concerns in using our findings to make sense of prior research on the side-effect effect. The first is that compared to previous demonstrations of the effect, the differences between the norm-conforming and norm-violating conditions in Experiment 1 are modest, and the "reverse" side-effect effect in the supervillain context from Experiment 2 is numerically smaller than that in the more typical, superhero context. A second potential concern is that while we find systematic differences in ascriptions of intentional action across our scenarios, it's not always the case that a majority of participants provide "intentional" ratings in the norm-violating cases (i.e. ratings above the scale midpoint) and a majority provide "unintentional" ratings in the norm-conforming cases (i.e. ratings below the scale midpoint), as has been found in the past for the CEO vignette, among others.

In evaluating these concerns it's important to note that our vignettes were designed such that the actions and outcomes were identical across scenarios that varied in norm status. The fact that *any* differences were observed across matched vignettes supports a role for norm status. Moreover, it's likely that norms other than those the vignettes manipulated influenced the absolute ascriptions of intentional action, if not the differences across matched cases. For example, in the CEO vignette from Experiment 1, participants presumably applied the norm that environmental harm is bad in both the norm-conforming and norm-violating conditions, generating ratings that were typically above the midpoint in both conditions. Finally, our vignettes required participants to accept a stipulated norm rather than employ their own norms, requiring non-trivial perspective taking. This is especially apparent in Experiment 2. It's impressive that norm status had a reliable effect above and beyond the effects of other norms that operated in the vignettes, participants' own norms, and additional factors that may contribute to ascriptions of intentional action.

Experiment 3

In Experiment 3, we turn to another prediction of the Rational Scientist view: that asymmetries in mental state ascriptions should track differences in predictions of future behavior. According to the Rational Scientist view, theory of mind serves the function of predicting and explaining behavior. It follows that mental state terms should track aspects of behavior that support prediction. Experiment 3 examines this aspect of the Rational Scientist view by considering whether *norm-violating* behavior, which supports a stronger ascription of intentional action than does *norm-conforming* behavior, also supports stronger predictions. More precisely, we suggest that background information supports mental state and trait inferences (MST1) that are updated in light of what an agent says and does (yielding MST2). In the case of norm-conforming behavior, MST2 will be very similar to MST1. In the case of norm-violating behavior, MST2 may differ substantially from MST1. If mental states and traits are posited to support predictions about future behavior, then norm-violating behavior should lead to predictions that deviate more from baseline predictions than does norm-conforming behavior.

To test these predictions, we consider three conditions. In the norm-conforming and norm-violating conditions, agents bring about good or bad side effects, respectively. In the *baseline* condition, agents do not perform actions or bring about side-effects. Then, in all conditions, instead of having participants judge whether a side effect was brought about intentionally, they make two predictions about the agent in the vignette's future behavior. The *specific* prediction considers whether the agent is more likely to engage in a norm-conforming or norm-violating behavior in the future. The *general* prediction concerns the agent's broader adherence to norms, and thus examines whether the inferred properties of the agent are restricted to the specific outcome in the vignette (e.g. harming the environment) or generalize more broadly (e.g. harming in general). The baseline condition should track the predictions supported by MST1; the norm-conforming and norm-violating conditions should track the predictions supported by MST2, where MST2 will differ across conditions in light of the agent's norm-conforming or norm-violating behavior.

The Rational Scientist view predicts that participants who learn about the agent who generates a norm-violating side-effect will make predictions about the agent's future behavior that differ more from baseline predictions than will participants who learn about the agent who generates a norm-conforming side-effect. In contrast, the Intuitive Moralist and Biased Scientist views focus primarily on the role of evaluative considerations in ascriptions of intentional action, and do not explicitly bear on the relationship between such ascriptions and predictions about future behavior. While the views could potentially be modified or supplemented to generate a prediction, they do not do so in their current forms.

Participants

Participants were 156 University of California-Berkeley undergraduates who participated for course credit.

Materials and Procedure

Participants were randomly assigned to one of three conditions: *baseline*, *norm-conforming*, or *norm-violating*. Participants in the *norm-conforming* and *norm-violating* conditions were presented with two short vignettes, the CEO vignette (Knobe, 2003a) from the introduction as well as the analogous DR vignette:

DR Vignette:

A team of doctors is treating a patient. One doctor on the team came to the senior doctor and said, "We are thinking of starting a treatment. It will lower the patient's blood pressure but it will also help [hurt] the patient's stomach problems."

The senior doctor answered, "Stomach problems are not our concern. I just want to lower the patient's blood pressure as much as I can. Let's start the treatment."

They started the treatment. Sure enough the patient's stomach problems were helped [hurt].

After each vignette participants were asked to make two ratings about the future actions of the agent in the story, a *specific* prediction and a *general* prediction. These questions are below, with the text for the CEO vignette in brackets:

Specific prediction:

In the following month the doctor [chairman] will make another decision that results in either:

A. An action that has a positive consequence beyond what the doctor is treating. [that helps the environment]

Or **B.** An action that has a negative consequence beyond what the doctor is treating. [that harms the environment]

Which decision do you think the doctor [chairman] will make?

General prediction:

The next month the doctor [chairman] will make another decision that results in either:

A. Exceeding ethical standards.

Or **B.** Violating ethical standards.

Which decision do you think the doctor [chairman] will make?

Participants rated the likelihood of each event on a scale from 1 to 7, where 1 indicated "very likely to choose A," 4 "equally likely to choose A or B," and 7 "very likely to choose B."

Participants in the *baseline* condition were introduced to the agents (e.g. "There is a chairman of the board who makes the final decisions for his company") and made all four prediction judgments, but were given no information about the agents' past behavior.

The order of story presentation (CEO first or DR first) and the direction of the 7-point scale (from conforming to violating or vice versa) were counterbalanced across participants.

Results and Discussion

To examine whether participants' prediction ratings varied across conditions, the data were first reverse-coded for participants who received a 7-point scale with higher

values indicating a greater probability of acting to bring about a positive side-effect. Thus for all participants, higher ratings correspond to a higher subjective probability that the agent will act to bring about a negative side effect. We then conducted an ANOVA with condition as a between-subjects variable (*baseline*, *norm-conforming*, *norm-violating*), vignette as a within-subjects variable (CEO, DR), prediction question as a within-subjects variable (specific, general), and prediction rating as the dependent variable. This revealed a main effect of condition ($F(2,153)=14.36$, $p<.001$), as well as a main effect of vignette ($F(1,153)=83.43$, $p<.001$). Overall, participants rated negative actions more probable in the *norm-conforming* condition than in the *baseline* condition, and in the *norm-violating* condition than in the *norm-conforming* condition (see Fig. 1.3). Ratings in the *norm-conforming* condition may have been more negative than in the *baseline* condition because failing to endorse a fortuitous side effect (e.g. helping the environment) is itself a norm violation (see Mele & Cushman, 2007). The main effect of vignette resulted from the fact that predictions concerning the CEO were generally more negative than those concerning the doctor.

The key hypothesis that predictions in the *norm-violating* condition should differ more from baseline than do those in the *norm-conforming* condition can be examined by looking for significant differences across these conditions, as both yielded ratings more negative than those in the *baseline* condition. An ANOVA like that above but restricted to the *norm-violating* and *norm-conforming* conditions reproduced the main effect of vignette ($F(1,102)=50.86$, $p<.001$) and revealed a main effect of condition ($F(1,102)=8.75$, $p<.01$) as well as a 3-way interaction between vignette, prediction, and condition ($F(1,102)=4.80$, $p<.05$). With post-hoc t-tests, the *norm-conforming* and *norm-violating* conditions differed significantly on both CEO predictions (specific: $t(102)=3.43$, $p<.001$; general: $t(102)=2.18$, $p<.05$), and were marginal for the general DR predictions (specific: $t(102)=1.11$, $p=.271$; general: $t(102)=1.91$, $p=.059$).[4] These findings confirm the prediction that relative to baseline, norm-violating behavior provides more information about an agent's future behavior than norm-conforming behavior.

Although our task did not require participants to report the mental states ascribed to the agents in each vignette, the nature of their predictions provides some evidence concerning these mental state ascriptions. Recall that participants made two kinds of predictions: a specific prediction about the same norm-violation in the future, and a general prediction about norm-violation in general. The fact that the predicted pattern of results was obtained for both kinds of predictions suggests that participants not only ascribed the agents in each vignette with a specific attitude concerning the violated norm (e.g. that the CEO does not value the environment or that the DR is insensitive to

[4] To verify that the DR vignette generates a side-effect effect, a different group of 72 participants was randomly assigned to either the CEO or the DR vignette in a condition involving either a helpful or a harmful side-effect. On a 7-point scale, participants judged whether it was appropriate to say that the agent *intentionally* brought about the side effect. This experiment revealed a main effect of condition ($F(1, 68) = 121.5$, $p < .001$) as well as an interaction between condition and vignette ($F(1, 68) = 9.82$, $p = .003$). The help/harm asymmetry was smaller for the DR (2.3 for help versus 4.5 for harm) than for the CEO (1.4 for help versus 5.3 for harm), but even the DR vignette involved a significant effect of condition ($t(34) = 5.13$, $p < .001$).

patients' overall well-being), but also ascribed the agents with a more general trait (e.g. the CEO is evil) or a general attitude towards norms (e.g. the DR thinks he can ignore the rules).

Of the views that have been proposed, only the Rational Scientist view provides an explanation for why norm-violating behavior would support stronger predictions than norm-conforming behavior. Accordingly, only the Rational Scientist view predicts the findings from Experiment 3. However, the Intuitive Moralist and Biased Scientist views could be modified to accommodate these findings. In particular, the Intuitive Moralist view could stipulate that the valence of an outcome influences mental state ascription in general (beyond ascriptions of intentional action), with consequences for prediction, and one of the most recent formulations (Pettit & Knobe, 2009) does extend beyond ascriptions of intentional action. Similarly, Biased Scientist models could build in a mechanism by which judgments of praise or blame bias all mental state ascriptions, which in turn influence predictions. So while the findings from Experiment 3 are specifically predicted by the Rational Scientist view, the greatest contribution of Experiment 3 may be to highlight the intimate relationship between mental state ascriptions and prediction.

General Discussion

The three studies presented suggest that norm status is sufficient to produce a side-effect effect, and that moral status is not necessary. In particular, the findings demonstrate that norm status can generate a side-effect effect when moral status is controlled (Experiment 1 & 2), that conventional norms can also generate a side-effect effect (Experiment 1), and that norm-violating behavior has a greater influence on future predictions than does norm-conforming behavior (Experiment 3). These findings are predicted by the Rational Scientist view, according to which norms influence mental state ascriptions because norm-violating behavior supports the ascription of counternormative mental states, which in turn influence ascriptions of intentional action, predictions of future behavior, and other judgments relevant to theory of mind.

According to the Rational Scientist view, mental states and traits (MST1) are ascribed to novel agents on the basis of context, norms, and other available information. After observing a behavior – such as a CEO denying an interest in the environment or proceeding with a risky plan – observers update ascribed mental states and traits (generating MST2), with the behavior's relationship to norms as a source of evidence concerning the agent's mental states. In particular, moral (and other prescriptive) norms provide a reason for behaving in accordance with the norm, so a behavior that deviates from the norm suggests the existence of a conflicting reason for action – one sufficiently strong to outweigh the reason to conform to the norm. Positing such conflicting reasons may involve mental state ascriptions (e.g. "dislikes the environment," "is evil") that in turn generate different judgments.

So while there does seem to be an influence of moral evaluation on theory of mind judgments, the relationship may be best described as *evidential*. That is, the status of a behavior with respect to norms provides *evidence* about underlying mental states, but norm status need not be constitutively tied to folk psychological concepts like 'intentional action'. Instead, the judgment that an outcome was or wasn't brought about "intentionally" is a function of the mental states and traits ascribed to the agent (MST2),

with information about the outcome and the agent's causal contribution to its occurrence also likely to play a role.

Why would the mental state ascriptions licensed by norm-violating behavior lead participants to judge that a side effect was brought about intentionally? One possibility is that participants ascribe the mental states required by Malle and Knobe's (1997) account of intentional action. According to this account, the folk concept of intentional action involves five components: desire, belief, intention, awareness, and skill. In Knobe's original CEO vignette and in those in the current experiment, the agents believe their actions will produce the outcome in question, they perform actions with this awareness, and they have the requisite skills. This leaves "desire" and "intent" as components of intentional action that are not explicitly specified by the vignette, but that participants may infer in the norm-violating case. In particular, instances of norm violation provide a relative ranking of what the agent values. When the CEO violates an environmental norm, for example, one can infer that he values (desires) profit more than he values the aspect of the environment that will be harmed. But in the norm-conforming condition there is no equivalent information about how the CEO values the environment relative to profits. While in both cases the agent expresses no concern for the side effect, the agent's actions provide unambiguous mental state information in the form of a relative value only when a norm is violated. It may be that the low relative value of the environment in norm-violating cases is sufficient to satisfy the "desire" and "intent" requirements of Malle and Knobe's (1997) account of intentional action, even if the agent does not actively desire that the environment be harmed.

Another possibility is that people's understanding of intentional action centers on choice, with an action judged intentional when there are alternative options apparent to the agent (James, 1890; Tolman, 1925; Miller et al., 1960). Along these lines, some have suggested that intent is particularly clear when the agent makes the "hard choice by following a previously nondominant alternative" (Fiske, 1989). Perhaps participants ascribe intent in cases of norm violation because they involve a clear (and dominant) alternative.

Relationship to previous accounts

While other accounts of the side-effect effect can be modified to accommodate our findings, the Rational Scientist view has the advantage of specifically predicting the observed pattern of results. Moreover, the Rational Scientist view can accommodate several cases in the literature that have proved difficult for other accounts of the side-effect effect. We briefly review these cases and alternative theories, and then consider the role of norms in theory of mind more broadly.

Most accounts of the side-effect effect have focused on the influence of moral valence (good or bad) or moral evaluation (blameworthiness or praiseworthiness) on judgments of intentional action (e.g. Knobe, 2003; Knobe, 2006; Nadelhoffer, 2004b; Wright & Bengson, 2009). However, subsequent studies using similar vignettes have produced examples that counter these accounts. For example, Phelan and Sarkissian (2008) generated vignettes for which side effects were judged intentional but neither bad nor blameworthy, as well as others for which side effects were *not* judged intentional despite being judged bad. In one case, participants evaluated vignettes (from Knobe and Mendlow, 2004) in which the president of a corporation maximized company-wide sales, but as a side effect either decreased sales in one particular division or increased the

prominence of one division relative to another. Most participants judged that the president had intentionally performed both side effects, but did not judge the side effects to be either bad or blameworthy. In a vignette demonstrating the opposite pattern, a city planner reluctantly decides to implement a plan that increases joblessness as a side effect of cleaning up pollution. Participants rated the side effect as bad, but did not endorse the claim that it was brought about intentionally.

These results are difficult to accommodate with an account that focuses exclusively on moral valence or responsibility. However, the Rational Scientist view can explain these results. Because information about mental states is inferred from norm violations, the Rational Scientist view does not require side effects to be bad or blameworthy, only to be norm-violating. In the context of a corporation, a president operates under a norm to improve the corporation. The fact that the president is willing to incur a cost in the form of decreased sales in one division provides evidence that there must be a compelling reason to engage in the action – one sufficiently strong to outweigh a standing reason to increase sales. In the language of the Rational Scientist view, the baseline MST1 says that the president wants sales in all divisions to increase or stay the same. As in the CEO vignette, the action tells us about relative value: that the value assigned to sales in that division is lower than that assigned to the principle aim, in this case maximizing company-wide sales. This is evidence that MST1 does not provide a satisfactory picture of the president's mental states, suggesting a change to MST2 is necessary. This evidence about relative value may in turn influence ascriptions of intentional action.

In the case of the city planner, there is extra information about the agent's mental state. The city planner is choosing between adhering to two conflicting norms, one to decrease joblessness and another to clean up pollution. The city planner states that he "feel[s] awful" about the side effect. Because participants are told about the city planner's attitude towards the side effect (and they have no reason to doubt what they are told), they have no need to infer a desire or other mental state that could support an ascription of intentional action. (For a similar point see Guglielmo and Malle, 2009a.)

Machery (2008) proposes an account of the side-effect effect called the trade-off hypothesis that does not involve moral valence or responsibility. In his studies, participants evaluated non-moral situations, such as one in which an agent orders the largest smoothie available and as a side effect either pays an extra dollar or receives a free cup. Most participants judged that the agent paid the extra dollar intentionally, but that he did not receive the free cup intentionally. Machery suggests that the extra dollar is conceptualized as a *cost* incurred as a means to a benefit, and that costs are considered intentional. Because the free cup is not a cost that trades-off with the benefit, it is not judged intentional. However, Mallon (2008) provides examples of the side-effect effect that offer *prima facie* evidence against the trade-off hypothesis. The key vignettes involve agents who would not consider a "bad" side effect a cost. In one case, a terrorist intends to harm Americans and as a side effect either hurts Australians or helps orphans. According to the terrorist both side effects are good, so neither is a cost incurred for a greater benefit. However, participants responded that harming Australians was intentional but helping orphans was not, which Mallon argued was evidence against the trade-off hypothesis, since participants were willing to call a bad side effect intentional even when the agent did not view it as a cost.

We see the trade-off hypothesis as similar in spirit to the Rational Scientist view, but the Rational Scientist view is more general and can more easily accommodate examples like Mallon's. Conceptualizing costs in terms of norms and norm-violation can help explain both what is considered a cost, and why a cost might be considered intentional. The fact that an agent is willing to incur a cost provides evidence that the agent has a reason to perform the action that is sufficiently strong to outweigh the cost — we can infer that according to the agent, the benefit outweighs the cost. Costs thus play a similar evidential role to norm-violations.

Given the similarities between the trade-off hypothesis and the Rational Scientist view, Mallon's "no tradeoff" terrorist cases pose a potential challenge. In particular, why don't the terrorist cases generate a side-effect effect reversal, as in the supervillain context from Experiment 2? First, because the Rational Scientist view suggests that key mental states and traits are inferred on the basis of norm violations, it's difficult to know how to evaluate the terrorist cases without explicit guidance on the norms with respect to which the agent is operating. Although the terrorist doesn't consider harming Australians to be a cost, taking this statement at face value requires participants to suspend their own norms – precisely what Experiment 2 attempts to accomplish with the supervillain cover story by being *very* explicit about the agent's norms. Even if participants succeed in considering the vignette from the perspective of the terrorist, participants may have reasonably inferred a reason to harm Australians that outweighed a universal norm such as "do not harm for no reason." In the supervillain context, we aimed to eliminate such background norms by stipulating that the supervillains are the badest of the bad, look for every opportunity to cause harm, and so on. In contrast, there is no norm against helping orphans, so the same asymmetry as in the CEO problem emerges. (A similar argument can be made for interpreting the results of the Nazi identification problem used in Knobe, 2007.)

Additionally, the terrorist case only presents one side of the 2x2 design used in our Experiment 2 (superhero or supervillain context x helpful or harmful side effect). Reducing or eliminating a trade-off for all or some participants should have reduced the asymmetry in the side-effect effect, but this reduction wouldn't be apparent without conditions featuring a typical agent (i.e. a non-terrorist context) for comparison. Finally, the terrorist case differs from our own supervillain cases in the agent's expressed attitude towards the side effect. The terrorist acknowledges that the side effect would be a good thing in both conditions; the agents in our supervillain context claim indifference, but operate amidst norms that would dictate a positive attitude towards bad side effects (such as harming Australians) and a negative attitude towards good side effects (such as helping orphans).

Other accounts of the side-effect effect have been offered, but most have the characteristics of the accounts we have considered: they invoke a notion like moral valence or moral responsibility, or they appeal to a more general (non-moral) notion of goodness and badness. Because the Rational Scientist view emphasizes the relationship between an action and norms, involves tracking mental states, and allows for multiple sources of predictive information, it is equipped to address the kinds of cases that have proved problematic for such accounts, and provides a more complete explanation of the side-effect effect.

More recently, some have offered accounts suggesting that the side-effect effect is multiply determined (Sloman, Fernbach and Ewing, 2010; see also Gugiellemo and Malle, 2009; Sripada, 2009 for views that emphasize other factors). While we have argued that the Rational Scientist view is sufficient to explain observed asymmetries in judgments of intentional action, it is certainly possible that the factors highlighted by these accounts play an additional role in generating judgments.

Norms in Theory of Mind

The Rational Scientist view preserves the traditional functions of ToM, prediction and explanation, though additional functions are certainly possible. However, the Rational Scientist view also emphasizes a role for information about norms in prediction and explanation (see also Wellman & Miller, 2006; Kalish 2006). Specifically, norms play a critical role in establishing baseline mental state and trait inferences (MST1), and in determining how observations influence subsequent mental state and trait inferences (MST2). In the absence of evidence that an agent has counternormative mental states or traits, norms may support prediction and explanation directly – without being mediated by explicit mental state attributions.

Developmental research has suggested that for children under the age of four, moral and conventional norms are an important basis for explaining and predicting behavior (Kalish, 1998). For example, young children predict that an agent will conform to a norm, even if the norm is unknown to the agent or conflicts with the agent's own preferences. However, older children and adults predict that when norms and preferences conflict, preferences will often win out (Kalish & Cornelius, 2007; Kalish & Shiverick, 2004). Even in adults, not all belief inferences are automatic (Apperly, 2006); it's possible that norms directly support many everyday predictions and explanations, with the corresponding mental state inferences drawn only as needed.

Recognizing a role for norms in mental state ascriptions raises a number of important questions. For example, is the influence of norms on mental state ascriptions restricted to prescriptive norms, such as the conventional and moral norms considered here? We suspect a similar relationship holds for statistical "norms" or generalizations. A behavior that violates a statistical norm is not 'expected', and hence provides information about the agent's underlying mental states that may lead to a change from MST1 to MST2. If most people conform to a norm to drink coffee black, for example, observing someone drink black coffee is relatively uninformative: the behavior could have been predicted from the statistical norm. On the other hand, observing an agent violate this norm by adding cream and sugar *is* informative: rather than ascribing default mental states, we can ascribe an atypical attitude towards coffee (see Lucas et. al., 2009, for a similar argument). As with prescriptive norms, this makes sense if the function of ToM is to track information that supports prediction and explanation.

A related question concerns the interactions between multiple norms. While many moral norms are also statistical norms, there may be cases in which norm-conformance is rare, placing a moral norm in conflict with a statistical norm. How are mental state ascriptions made under such conditions? These cases may be uncommon because a moral norm would presumably be the statistical norm unless conformance had a cost. But as an illustrative example, consider the low-cost behavior of agreeing to donate one's organs in case of accidental death. Though it is generally believed that organ donation is morally good (*morally norm-conforming*), actual organ donor rates in the US are not very high

(*statistically norm-violating*) (Sheehy et. al., 2003). In this case, it may be possible to see a reversal of the typical side-effect effect, where the morally good behavior (organ donation) is more informative and judged intentional.

Conclusion

While we've contested Knobe's (2003, 2006) interpretations of the side-effect effect as a challenge to the traditional functions of theory of mind, our findings support the underlying claim that moral (and other) norms influence mental state ascriptions. The key lesson from our arguments and findings is that sensitivity to norms is central to the ability to predict and explain behavior.

Figures

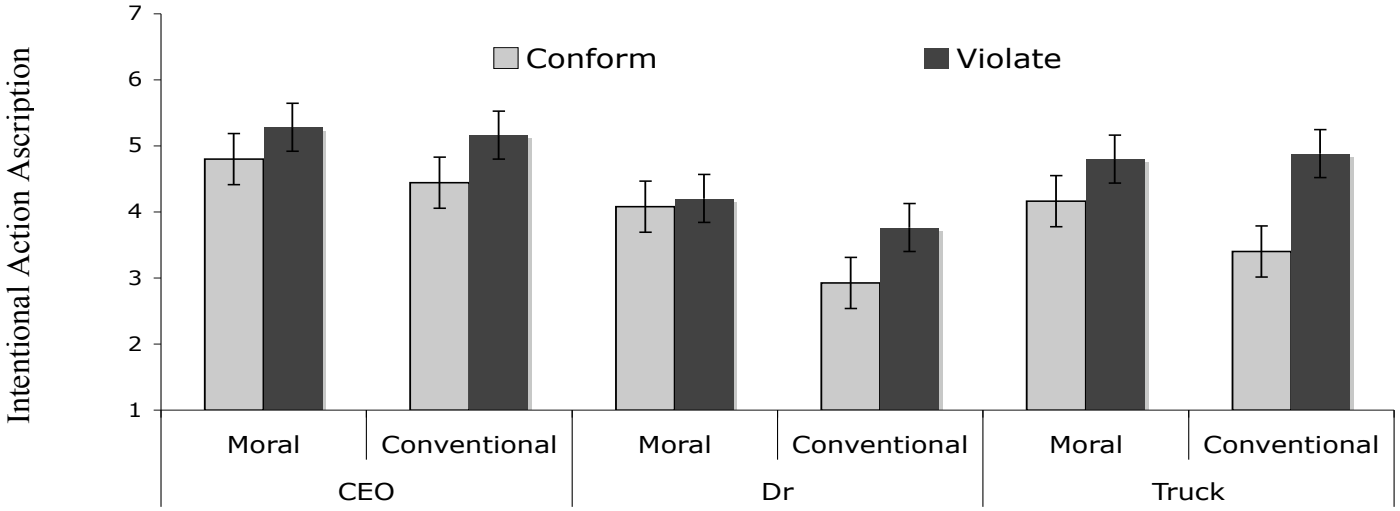


Figure 1.1 - Experiment 1 ratings of how appropriate it is to call an action intentional as a function of norm status and norm type. Ratings were made on a scale from 1 (neither appropriate nor inappropriate to say outcome brought about intentionally) to 7 (appropriate to say outcome brought about intentionally) with 4 (neither appropriate nor inappropriate) as a midpoint.

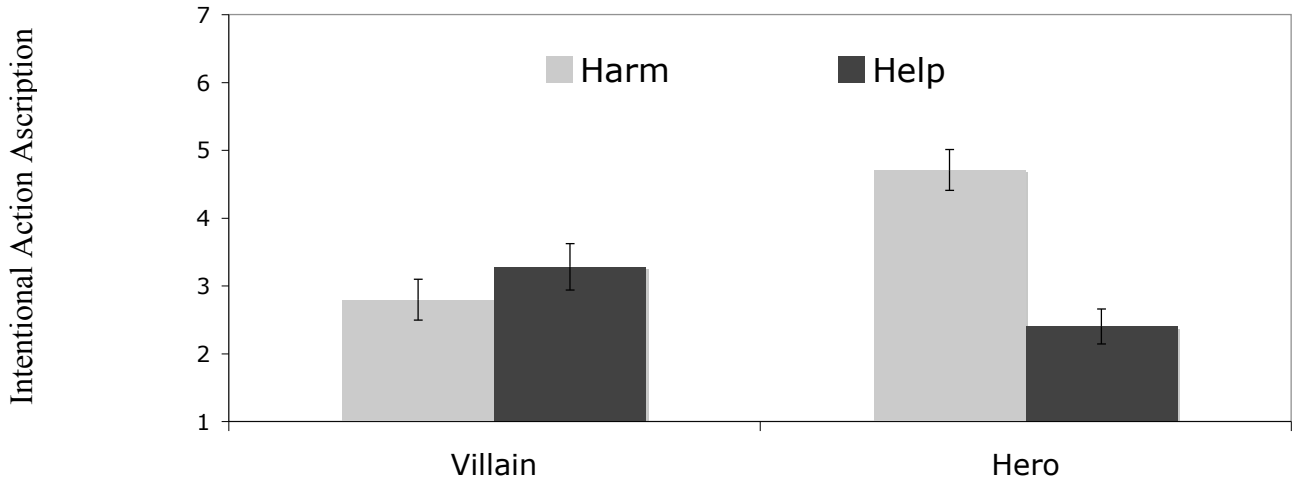


Figure 1.2 Experiment 2 ratings of how appropriate it is to call an action intentional as a function of norm status and context. Presented on a scale from 1 (not appropriate to say outcome brought about intentionally) to 7 (appropriate to say outcome brought about intentionally), with 4 (neither appropriate nor inappropriate) as a midpoint.

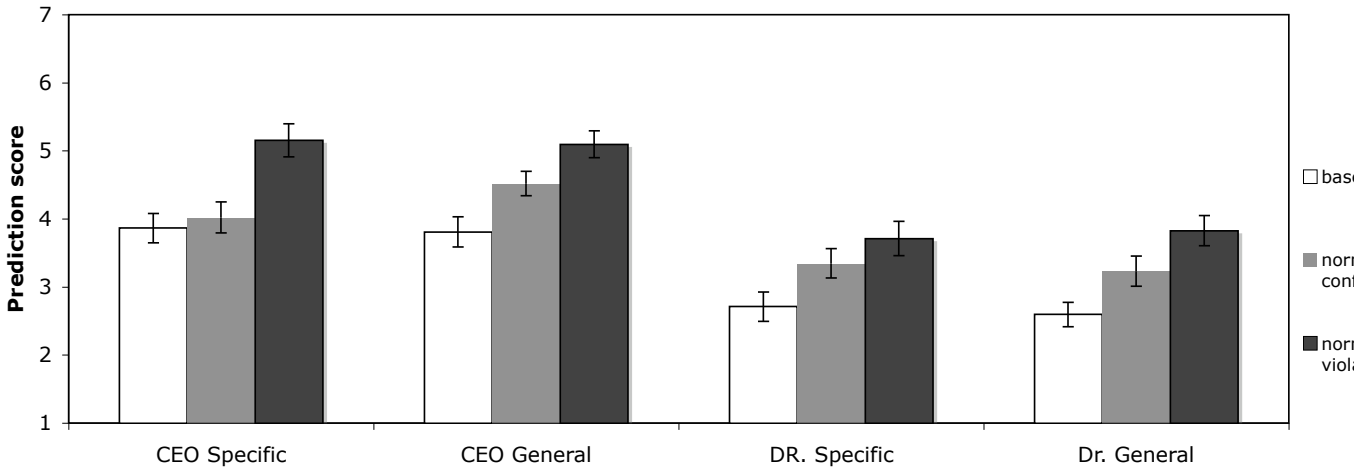


Figure 1.3 - Prediction scores from Experiment 3 on a scale from 1 (good side effect likely in future) to 7 (bad side effect likely in future).

Tables

Question	Superhero		Supervillain		Main effects		Interactions		
	Good SE	Bad SE	Good SE	Bad SE	SE valence	Context	SE valence x Context	Context x vignette	3-way
Questions from superhero or supervillain perspective:									
(a) Which do you think Bob is more likely to do in the future, [good or bad SE]? (7 = likely to [good SE])	4.08 (1.31)	3.67 (1.44)	3.17 (1.66)	3.12 (1.42)		*			
(b) Which do you think Bob is more likely to do in the future, [good or bad main effect]? (7= likely to [good ME])	5.71 (1.20)	5.88 (1.30)	1.42 (.88)	1.79 (1.25)		*		*	
(c) In comparison to Steve [an average candidate], do you think Bob is more or less likely to [bad SE]? (7 = more likely)	4.17 (1.27)	4.83 (1.09)	4.46 (1.50)	4.17 (1.01)					*
(d) In comparison to Steve, do you think Bob is more or less likely to [good SE]? (7 = more likely)	4.21 (1.38)	3.50 (1.18)	3.46 (1.02)	4.25 (1.03)			**		
(e) Was [SE] good or bad? (7 = definitely good)	6.33 (1.05)	2.42 (1.10)	4.04 (2.26)	4.29 (2.49)	**		**		
(f) Is Bob a good or bad *person*? (7 = definitely good)	4.83 (1.34)	4.12 (1.26)	1.87 (.74)	2.37 (1.35)		**	*		**
(g) Should Bob be blamed or praised? (7 = praised)	4.92 (1.21)	3.29 (1.04)	4.25 (1.51)	4.83 (1.71)			**		
(h) How does observing Bob's effect on [SE] impact your recommendation to the Supervillain? (7 = recommend promotion)	3.37 (1.28)	2.83 (1.71)	3.04 (1.27)	4.08 (1.79)			*		
Questions from personal perspective:									
(j) Do YOU think [SE] was good or bad? (7 = good)	6.29 (1.08)	2.17 (1.24)	6.21 (1.25)	1.62 (1.25)	**				
(j) Do YOU think Bob is a good or bad *person*? (7 = good)	4.67 (1.20)	3.96 (1.12)	1.96 (.96)	1.83 (1.13)		**			*
(k) Do YOU think Bob should be blamed or praised? (7 = praised)	4.79 (1.29)	3.12 (.99)	2.67 (1.66)	1.79 (1.14)	**	**			*

* $p < .05$, ** $p < .01$. Standard deviations follow means. Shading highlights significant effects.

Table 1.1 - Judgments from Experiment 2 as a function of context and side-effect valence. Means are followed in parentheses by standard deviations.

Chapter 3: Norms as explanations for behavior.[5]

In chapter three we investigate the use of norms as explanations for the behavior of an agent as well as the impact of norm and belief information on other explanations of behavior. We examine whether the use of norms in the social-cognitive capacity of explanation is simply shorthand for a causal process that includes an explicit belief or whether explicit belief is not the only acceptable causally explanatory pathway

Introduction

When explaining the behavior of others, people often appeal to beliefs, desires, and other mental states. For example, one might explain why Alan ordered a martini by mentioning his desire for gin, or explain why Barbara opened a box by mentioning her belief that chocolate was inside. However, some explanations for behavior appear to bypass mental states altogether. For example, imagine Carl setting a place at a table and putting the fork to the left of the plate. One could explain Carl's behavior by appeal to his desires or beliefs, with an explanation such as, "he *wanted* to place the fork to the left of the plate" or "he *believes* that the fork goes there." But in this example, and in many cases like it, it seems more natural to explain the behavior by (implicitly) referencing a social norm: "he put the fork to the left of the plate because that's where it is supposed to go." We refer to such explanations as "norm explanations." What do norm explanations reveal about social cognitive reasoning? More specifically, what do they reveal about the way that people understand the relationship between norms, mental states, and behavior?

One way to make sense of norm explanations is to assume that they are shorthand for a causal process that includes mental states (see Fig. 2.1). So, for example, the norm explanation offered above could assume that beliefs and desires mediate the relationship between the norm and behavior, such as the agent's belief that the norm is to place the fork to the left and his desire to conform to the norm. Consistent with this idea, research suggests that even young children appreciate the mediating role of mental states when it comes to following social rules (Kalish, 1998). For example, most 5-year-olds will judge that Arnie, who "doesn't know that kids don't wear shoes in the bathtub" and "wants to wear shoes in the bathtub" will wear shoes in the bathtub, but realize that a little boy who doesn't know that he can't turn into a bird and fly will not succeed in doing so, even if he wants to (Kalish, 1998, study 3).

On the other hand, there is evidence to suggest that reasoning about behavior does not necessarily and automatically involve explicit recognition of intervening mental states. For example Apperly, Riggs, Simpson, Chiavarino, and Samson (2006) presented participants with a task that sometimes required them to report an agent's beliefs, and found that participants were slower to answer such questions when they were not specifically instructed to track belief in advance. However, there was no equivalent processing cost for questions concerning the state of the world (in this case, an object's true location), suggesting that while tracking the state of the world is automatic, agents' beliefs are inferred only as necessary. Relatedly, Lin, Keysar, and Epley (2010) found that participants with lower working memory or who completed an attentionally-

[5] The material in chapter 3 was co-authored with Dr. Tania Lombrozo.

demanding secondary task were less effective at considering another person's intentions, with no equivalent cost for reasoning on the basis of their own beliefs.

The findings from Apperly et al. (2006) and from Lin et al. (2010) suggest that at least some reasoning about agents and behavior occurs in the absence of explicitly posited mental states. Instead, people seem to rely on sources of evidence that are more readily available than (unobserved) mental states: the true state of the world in Apperly et al. (2006), and one's own beliefs in Lin et al. (2010). Norms could potentially play a similar role, supporting many everyday judgments in the absence of an explicitly articulated mechanism by which mental states mediate the relationship between norms and behavior. In other words, people could reason as if norms provide direct explanations for behavior, only positing and articulating intervening mental states as necessary.

In this paper we investigate the relationship between norms, behavior, and mental states by considering the conditions under which people accept and produce norm explanations for behavior. In particular, we consider whether people accept and produce norm explanations when the relevant norm does or does not match the behavior, and when it is or is not accompanied by an agent's correct belief about the norm. To illustrate, consider a case in which the norm is to place the fork to the left of the plate and an agent sets a table by placing the fork to the left (matching the norm), but *believes* that the norm is to place the fork on the right (mismatching both the norm and behavior). Is it appropriate to explain the behavior by direct appeal to the norm, explaining, for example, that "the agent placed the fork on the left because the norm is to place it there"? If people regard a belief in the appropriate norm as a necessary causal link between the norm and behavior, then such explanations should be rejected. In the two experiments that follow, however, we show that such explanations are accepted and produced fairly often, and argue that norm explanations may not require a fully worked out causal pathway between norms and behavior.

A second question we consider is whether different kinds of norms play equivalent roles in explaining behavior. We examine three types of norms: moral (e.g., one should not resolve conflicts with violence), conventional (e.g., the fork goes to the left of the plate), and statistical (e.g., people tend to wear black shoes). While moral and conventional norms are both prescriptive – and therefore specify how someone *ought* to behave – statistical norms are purely descriptive. Moral and conventional norms thus provide desire-independent reasons for action (Searle 2001, see also Uttich & Lombrozo, 2010): An agent has a reason to conform to moral and conventional norms whether or not she wants to, and whether or not she ultimately does. As a result, norm explanations could be more strongly favored or more frequently produced when moral or conventional norms are involved than when only statistical norms are involved.

Moral norms have additional properties that set them apart from both conventional and statistical norms: They are often regarded as universal (Turiel, 1983; but see Kelly, Stich, Haley, Eng, & Fessler, 2007), and they are often regarded as objective in the sense that they do not depend on an individual's or group's beliefs (Goodwin & Darley, 2008; but see Sarkissian, Park, Tien, Wright, & Knobe, 2011). If moral norms are relatively "belief-independent," then norm explanations involving moral norms could be regarded as more explanatory than other types of norms even when they are not supported by corresponding beliefs.

We examine these issues in two experiments. In both experiments participants read a short vignette that included information about a norm, an agent's belief about the norm, and an agent's action. In addition to varying the type of norm (moral, conventional, or statistical), we also manipulated the agent's action (e.g., placing the fork to the left or the right), whether the agent's belief about the norm matched the action, and whether the norm matched the action. Participants were then asked to explain the agent's behavior and either rated provided explanations (Experiment 1) or provided free responses (Experiment 2). This design allowed us to examine whether and how norm explanations depend upon the presence of corresponding beliefs, and whether and how this relationship varies across different types of norms.

Experiment 1

Experiment 1 investigated whether people accept explanations for behavior that appeal to norms, where the norms do or do not match an agent's behavior, do or do not match the agent's beliefs about the norm, and are moral, conventional, or statistical. The experiment additionally varied whether the norms were typical for our participants (e.g., putting the fork to the left) or atypical (e.g., putting the fork to the right), and examined ratings for explanations that appealed to an agent's belief (e.g. "he placed the fork there because he believes it is the norm to do so") or preference (e.g., "he placed the fork there because that's where he prefers it").

If participants believe that norms can only explain behavior when mediated by correct beliefs about the norm, then norm explanations should only be accepted when both the norm and the belief match the agent's behavior. Critically, norm explanations should not be accepted when the norm matches the agent's action but the agent's belief does not (e.g., if the norm is to place the fork on the left, the agent believes the norm is to place it on the *right*, and the agent in fact places it on the left). If, in contrast, norm-explanations are independent of this belief, and are believed to influence behavior by an alternative or unspecified causal route, then participants could still be willing to explain a belief-inconsistent action by appeal to the norm (e.g., "he placed the fork on the left because that is the norm").

Additional predictions concern differences across norm types. If participants treat prescriptive norms as more appropriate explanations than descriptive norms, then they should rate norm explanations more highly for moral and conventional norms than for statistical norms. And if moral norms are treated as more objective and belief-independent than other kinds of norms, then participants could be especially insensitive to the presence of a corresponding belief in the norm when it comes to evaluating norm explanations for moral norms. In other words, the quality of norm explanations could be less dependent on the presence of a matching belief.

Experiment 1 also allowed us to examine a host of interesting but secondary issues, including the conditions under which participants infer preferences as explanations and whether this varies as a function of the kind of norm involved. Based on previous work, we predicted that participants will be especially inclined to infer preferences for norm-violating actions (Uttich & Lombrozo, 2010).

Participants.

Participants were 768 University of California-Berkeley undergraduates who received the questionnaire as part of a larger packet completed for course credit. There were 48 conditions with 16 participants in each condition.

Materials and procedure.

Participants read a short vignette that introduced a novel group and an agent who was a member of the group. The vignette also included information about a norm that was held by the group, indicated what the agent believed the norm to be, and described a norm-related action that the agent performed. The norm was either moral (involving harm), conventional (involving social or cultural practices), or statistical (involving descriptive regularities). Below is a sample vignette involving a conventional norm:

> In the "Toma" culture there are norms about how to set a table. In particular it is the norm to place the fork to the left [right] of the plate.
> Bob is a member of this relatively unknown group. Bob believes that it is the norm to place the fork to the left [right].
> One day last week when he set the table Bob placed the fork to the left [right] of the plate.

After reading the vignette, participants were asked to rate explanations. For the sample vignette above, they saw the following question and candidate explanations in the order presented [labels in brackets were not included for participants]:

> "Based on what you know or can infer about Bob, Please rate the explanations for the following question:
> "Why did Bob place the fork to the left of the plate?"
> [Preference explanation] Because Bob prefers to place it there.
> [Belief explanation] Because Bob believes the norm is to place it there.
> [Norm explanation] Because the norm is to place it there.

Explanations were rated on a 1 to 7 scale, with 1labeled "poor explanation" and 7 labeled "good explanation." Additionally, participants answered the following questions about the agent and their own beliefs about the norms, with responses made on a 1 to 7 scale [with response labels in brackets]:

> How strong is Bob's preference?
> [1 "Not at all strong" – 7 "Very strong"]
> How strongly does Bob believe this *should or shouldn't* be the norm?
> [1 "Shouldn't be the norm" – 7 "Should be the norm"]
> How strongly do YOU believe this *should or shouldn't* be the norm?
> [1 "Shouldn't be the norm" – 7 "Should be the norm"]
> How strongly do YOU believe this IS the norm?
> [1 "Certainly IS NOT the norm" – 7 "Certainly IS the norm"]

There were two norm topics for each of the three norm types (moral, conventional or statistical), resulting in six distinct sets of vignettes (See Table 2.1 for list of norms). The vignettes also varied according to three additional between-subjects' factors: norm

typicality, norm-action agreement, and belief-action agreement. Norm typicality refers to whether the norm was typical (e.g., fork on the left), or atypical, reversing the familiar norms for our participants (e.g., fork on the right). Norm-action agreement and belief-action agreement refer to whether the agent's behavior matched the group's norm or the agent's belief, respectively. For example, in the sample vignette above the agent's behavior ("Bob placed the fork to the left") matched both the group norm ("it is the norm to place the fork to the left") and the agent's belief ("Bob believes that it is the norm to place the fork to the left"). However, in alternative versions of the vignette these could mismatch, for a total of eight different combinations (right/left for behavior, right/left for group norm, right/left for agent's belief; see Table 2.2 for an example of possible norm, belief, and action combinations).

With the variables of norm type (3: moral, conventional, statistical), norm typicality (2: typical, atypical), norm-action agreement (2: match, mismatch), and norm-belief agreement (2: match, mismatch), plus the two versions of each norm type, there were 48 distinct vignettes, with participants randomly assigned to a single vignette.

Results and Discussion

Before moving on to our central results, we confirm our assumption that the typical norms were in fact typical for our participants. To do so, we examined whether responses to "How strongly do YOU believe this *should or shouldn't* be the norm?" and "How strongly do YOU believe this IS the norm?" as a function of norm typicality. A t-test revealed significant differences in the predicted direction for both this first question (typical M = 3.82, atypical M = 3.29, t(766) = 4.591, p < .01) and the second (typical M = 4.03, atypical M = 3.69, t(766) = 2.808, p < .01), although effects were not uniform across norm types.[6]

Below we present the results of analyses for each explanation type: Norm, Belief, and Preference. For each explanation type, we performed a 2 x 2 x 2 x 3 ANOVA with norm-action agreement (2: match, mismatch), belief-action agreement (2: match, mismatch), norm typicality (2: typical, atypical), and norm type (3: moral, conventional, statistical) as between-subjects factors and participants' ratings for the corresponding explanation as the dependent measure.

Norm explanation ratings. Norm explanation ratings are presented in Figure 2.2. Not surprisingly, norm explanations received higher ratings when the norm in the explanation matched the agent's action than when the norm did not match the agent's action (M = 4.37, SD = 2.02 versus M = 2.18, SD = 1.60), $F(1,768) = 282.131$, $p < .01$.

[6] T-tests were also performed with responses split by norm type. These analyses revealed significant differences between typical and atypical norms for both questions when the norm was moral "YOU believe this *should or shouldn't*?" (typical M = 4.27, atypical M = 2.82, $t(254) = 5.992$, $p < .01$) and "YOU believe this IS?" (typical M = 4.48, atypical 3.75, $t(254) = 3.730$, $p < .01$). When the norm was conventional this analysis revealed a significant difference for "YOU believe this *should or shouldn't*?" (typical M = 4.06, atypical M = 3.73, $t(254) = 2.098$, $p = .037$) but not "YOU believe this IS?" (typical M = 4.16, atypical 3.87, $t(254) = 1.330$, $p = .185$.). Differences between typical and atypical norms were not significant for statistical norms: "YOU believe this *should or shouldn't*?" (typical M = 3.13, atypical M = 3.32, $t(254) = -1.187$, $p = .236$) and "YOU believe this IS?" (typical M = 3.45, atypical 3.45, $t(254) = 0.37$, $p = .970$.)

Norm explanations also received higher ratings when the agent's belief about the norm matched the action of the agent than when the agent's belief mismatched the action (M = 3.48, SD = 2.11 versus M = 3.08, SD = 2.13), $F(1, 768) = 9.238$, $p = .002$. So, for example, explaining Bob's action of placing the fork on the left was better explained by the norm ("Because the norm is to place it there") when Bob believed that the norm was to place it there. However, the causal model in Figure 2.1 would predict an interaction between norm-agreement and belief-agreement (i.e., if norm explanations were only accepted when both the norm and the belief matched the action, as in the N+B+ condition), and this was not observed, $F(1, 768) = .014$, $p = .91$. Explanation ratings in the critical case involving a norm that matched the action and a belief that mismatched the action (N+B-) were still quite high, suggesting that while norm explanations are sensitive to corresponding beliefs, such beliefs are not judged necessary by all participants.

There was also an effect of norm type, $F(1,768)= 3.907$, $p = .021$, with norm explanations receiving higher ratings when the norm was conventional than moral or statistical (conventional: M = 3.51, SD = 2.19; moral: M = 3.27, SD = 2.09; statistical: M = 3.06, SD = 2.08). This could reflect a genuine difference across domains, with norm explanations judged more natural in the conventional case, or simply be an artifact of our particular stimulus materials.

There were no additional significant effects

Belief explanation ratings. Belief explanation ratings are presented in Figure 2.2. Not surprisingly, belief explanations received higher ratings when the agent's belief about the norm matched the agent's action (M = 5.51, SD = 1.71 versus M = 2.68, SD = 2.01), $F(1,768) = 485.948$, $p < .01$. However, this effect was not uniform across different types of norms (belief-action agreement x norm type interaction), $F(2, 768) = 9.791$, $p < .01$: The difference between belief explanation ratings across the matching and mismatching cases was smallest for moral norms, with a difference of 2.02, versus differences of 3.22 for conventional norms and 3.23 for statistical norms. In other words, participants' judgments about belief explanations were less sensitive to the actual match between belief and actions in the moral case.

Belief explanations were also rated more highly when the norm of the group matched the agent's action (M = 4.58, SD = 2.22 versus M = 3.61, SD = 2.35), $F(1, 768)= 57.432$, $p < .01$. So, for example, explaining Bob's action of placing the fork on the left was better explained by the belief ("Because Bob believes the norm is to place it there") when the norm in fact matched the action.

There were no other significant effects.

Preference explanation ratings. Preference explanation ratings are presented in Figure 2.2. Preference explanations were rated more highly when the norm of the group did *not* match the agent's actions (M = 2.84, SD = 1.73 versus M = 3.27, SD = 1.86), $F(1, 768) = 12.158$, $p = .001$. This could be because participants inferred that an agent who violated a norm must have had a reason for doing so, such as a preference for the norm-violating outcome (see Uttich & Lombrozo, 2010).

There was also a main effect of norm type, $F(2, 768) = 27.083$, $p < .01$, with preference explanations receiving lower ratings for conventional than either moral or statistical norms (Conventional: M = 2.45, SD = 1.66; Moral: M = 3.18, SD = 1.76; Statistical: M = 3.54, SD = 1.83). One possibility is that conventional norms are more

likely to be regarded as arbitrary decisions between equivalent options, and therefore less likely to be supported by strong, individual preferences. However, this effect could also be an artifact of the particular materials in our experiment.

Finally, there was an interaction between belief-action agreement and norm type, $F(2, 768)= 11.490$, $p < .01$. For moral norms, preference explanations were rated more highly when the agent's belief about the norm matched the agent's action than when they mismatched. However, for conventional and statistical norms, preference explanations were rated more highly when the agent's belief about the norm did not match the agent's action. Perhaps in the moral domain, participants believe that agents prefer what they believe or believe what they prefer, whereas for the other domains, preferences are more likely to be inferred when a belief explanation is inadequate.

Correlations between explanation ratings and other measures. To further understand why participants provided particular explanations, the ratings for each explanation type were correlated with responses to the questions concerning the strength of the agent's preferences, the agent's belief about what the norm should be, the participant's belief about what the norm should be, and the participant's belief about what the norm actually is.

Norm explanation ratings correlated negatively with ratings for the strength of the agent's preference ($r = -.221$, $p < .001$) but positively with ratings for the strength of the agent's belief about what the norm should be ($r = .224$, $p < .001$). In other words, and consistent with our previous results, participants were more likely to accept norm explanations when they did not believe that the agent had a strong preference, but when they did believe that the agent had a supporting belief. There was also a significant relationship between norm explanation ratings and ratings for the strength of the participant's own belief about what the norm should be ($r = .080$, $p = .026$) and belief about what the norm is ($r = .188$, $p < .001$), although both relationships were small.

Not surprisingly, belief explanation ratings correlated with ratings for the agent's belief about what the norm should be ($r = .364$, $p < .001$). They were also positively related to ratings for participants' own beliefs about what the norm should be ($r = .083$, $p = .021$), although once again the relationship was small.

Finally, preference explanations were significantly correlated with ratings for the strength of the agent's preferences ($r = .383$, $p < .001$).

There were no other significant correlations. These correlations, while generally small, are all in the expected directions and reinforce our previous results.

Summary of key results. Experiment 1 reveals that participants not only accept norms as explanations, but that the presence of a matching or mismatching norm can also impact the quality of belief and preference explanations. In particular, belief explanations were rated more highly when they matched the operating norm, while preference explanations were rated more highly when the action being explained was norm violating.

Interestingly, the findings provide only lukewarm support for the view that norms can explain behavior, but only when mediated by the agent's (correct) belief about the norm (Figure 2.1). While there was a statistically significant impact of beliefs on norm explanations, the effect was relatively weak, with no interaction between norm-agreement and belief-agreement, as would be expected if a matching norm and belief were required for an adequate norm explanation. Participants were surprisingly willing to endorse a

norm explanation when the norm matched the action being explained, even in the absence of a corresponding belief in the norm (see the N+B- condition in Figure 2.2). We return to this issue in motivating Experiment 2.

The findings do provide some positive support for the idea that different kinds of norms can play different explanatory roles. Overall, conventional norm explanations were rated more highly than moral or statistical norm explanations, which could reflect something particular to conventional norms rather than prescriptive norms more broadly. Perhaps more intriguing, there was support for the idea that moral norm explanations are less tied to mediating beliefs than other kinds of explanations. While norm explanation ratings did not interact with norm type, belief explanation ratings did, with a relationship in the predicted direction: The match between a belief and an action was less influential in the moral case than in the other cases.

Experiment 2

Experiment 1 established that norms matter for participants' endorsements of explanations, and additionally confirmed many of the predictions motivated in the introduction. However, the findings do not readily provide an explanation for why participants were so willing to accept norm explanations in the absence of a corresponding belief. While it's possible that participants believe norms can explain behavior even when they are not mediated by mental states, just as gravity can explain a person's fall independently of the person's mental states, this possibility is unlikely (Kalish, 1998). Instead, it could be that participants accept norm explanations in such cases because they either imagine an alternative causal route via which the norm could influence behavior (for example, through unconscious beliefs or unreflective imitation), or because their causal model is sufficiently underspecified that they are relatively insensitive to the absence of a key causal step (Keil, 2003). Experiment 2 addresses these possibilities by following the same basic design as Experiment 1, but requesting free responses to the explanation prompt rather than having participants rate provided explanations. This allows us to examine whether participants spontaneously provide norm explanations, and whether they invoke any alternative causal routes, such as unconscious beliefs or unreflective imitation.

Experiment 2 also provides an additional opportunity to examine whether prescriptive norms make for better explanations than descriptive norms, and whether moral norm explanations are less dependent upon a corresponding belief than other types of norms.

Participants.

Participants were 576 University of California-Berkeley undergraduates who received the questionnaire as part of a larger packet completed for course credit. There were 48 conditions with 12 participants in each condition.

Procedure.

As in Experiment 1, participants read a short vignette that introduced a novel group and an agent who was a member of the group. They were told a norm, the agent's belief about the norm, and the agent's norm-related action. Experiment 2 employed the same 48 distinct vignettes as Experiment 1, with participants randomly assigned to a single vignette. However, in Experiment 2 participants were prompted to provide explanations instead of rating provided explanations. Below is a sample prompt:

"Based on what you know or can infer about Bob, please explain the following in a sentence or two: Why did Bob place the fork to the left of the plate?"

In addition to their free response explanations, participants responded to the same questions about the agent's and their own attitudes towards the norm as in Experiment 1 (e.g., "How strong is Bob's preference?")

Results and Discussion

Explanation coding and analyses. Participants' explanations were coded for the presence of the three explanation types from Experiment 1: norm, belief (about the norm), and preference. A fourth explanation category included "alternative causal history" explanations (see Table 2.3 for samples from each category). To be coded as a norm explanation, an explanation had to explicitly cite the norm or the agent's membership in the group that held the norm. To be coded as a belief explanation, an explanation had to cite the agent's belief about the norm. Preference explanations had to cite an agent's preference, desire, or other positive attitude towards the action or outcome. Finally, alternative causal history explanations cited factors that impacted Bob and his action, but did not directly cite his belief about the norm. For example, this category included references to his unconscious or subconscious, peer pressure or social influence, and learning from teaching or observation. Occasionally, responses included multiple explanations of different types; these explanations were coded in multiple categories.

Agreement between two independent coders was 89% (Kappa = .688, $p < .001$). Disagreements were resolved by discussion.

We analyzed the frequency with which each explanation type was produced as a function of the same factors considered in Experiment 1. Specifically, we conducted log-linear analyses with the following factors: production of each explanation type (produced, not produced), norm-action agreement (match, mismatch), belief-action agreement (match, mismatch), norm type (moral, conventional, statistical), and norm typicality (typical, atypical). The results of these analyses are reported below for each explanation type.

Norm explanations. The frequency of norm explanations produced across conditions is presented in Figure 2.3. Mirroring the norm explanation ratings from Experiment 1, norm explanations were more likely to be produced when the norm matched the action, $c^2(1) = 85.758$, $p < .001$, and when the belief matched the action, $c^2(1) = 20.052$, $p < .001$.

In addition, there were two higher-order interactions. First, there was a three-way interaction between production of a norm explanation, belief-action agreement, and norm type, c^2 (2) = 17.124, $p < .001$. For both conventional and statistical norms, norm explanations were more often produced when the agent's belief matched the action. However, this relationship was not apparent for moral norms. Second, there was a three-way interaction between production of a norm explanation, belief-action agreement, and norm typicality, $c^2(2) = 4.175$, $p = .041$. Belief-action agreement had a larger impact on the frequency with which norm explanations were produced for typical norms than for atypical norms.

Mirroring Experiment 1, these findings suggest that norm explanations depend on a mediating belief, but not as strongly as might be expected. Participants were still quite willing to offer norm explanations in the absence of a mediating belief (see the N+B- condition in Figure 2.3). This was largely driven by the moral case, in which the frequency of norm explanations was independent of the presence of a corresponding belief.

Belief explanations. The frequency of belief explanations produced across conditions is presented in Figure 2.3. Not surprisingly, belief explanations were produced more frequently when the belief matched the action, $c^2(1) = 63.667$, $p < .001$. However, this effect was qualified by two higher-order interactions. First, there was an interaction between production of a belief explanation, belief-action agreement, and norm-action agreement, $c^2(2) = 10.818$, $p = .001$. When the agent's belief matched the action, belief explanations were more frequent when the norm mismatched the action (N = 58) than when the norm also matched the action (N = 37), potentially because the matching belief was more likely to be judged the only viable explanation. However, when the agent's belief did not match the action, belief explanations were more common when the norm did match the action (N = 15) than when the norm also mismatched (N = 6). The second interaction was between production of a belief explanation, belief-action agreement, and norm typicality, $c^2(2) = 3.938$, $p = .047$. Belief-action agreement had a smaller impact on the frequency with which belief explanations were produced for typical norms than for atypical norms.

Whereas Experiment 1 found that belief explanations were rated more highly in the presence of a matching norm, Experiment 2 found that belief explanations were more frequent in the absence of a matching norm. We suspect that this difference is a consequence of the different dependent measures. In Experiment 1, participants were asked to evaluate *each* explanation type, and could provide high ratings for more than one kind of explanation. In Experiment 2, participants might have stopped after producing what they took to be a single, sufficient explanation, and therefore failed to generate belief explanations when a norm explanation seemed more appropriate (i.e., in the N+B+ condition).

Preference explanations. The frequency of preference explanations produced across conditions is presented in Figure 2.3. Preference explanations were more likely to be produced when the norm did not match the action, $c^2(1) = 13.918$, $p < .001$, and when the agent's belief about the norm did not match the action $c^2(1) = 7.304$, $p = .007$, presumably because a preference was more likely to be inferred under these conditions. Preference explanations also varied across norm type, with a higher frequency for statistical norms than for the other norm types, $c^2(2) = 35.833$, $p < .001$.

These effects were qualified by two additional interactions. The first was between production of a preference explanation, belief-action agreement, and norm type, $c^2(2) = 11.041$, $p = .004$. For moral norms, preference explanations were more common when the agent's belief and action matched, with the reverse pattern for conventional and statistical norms. Second, there was a three-way interaction between production of a preference explanation, norm type, and norm typicality, $c^2(2) = 7.350$, $p = .025$. For moral norms, preference explanations were more common for atypical than for typical norms (19 versus 7); for conventional norms, preference explanations were about equally common for typical and atypical norms (16 versus 18); and for statistical norms, preference

explanations were more common for typical norms (39 versus 33).

Mirroring Experiment 1, Experiment 2 found, more preference explanations when the norm and action mismatched, and also – for moral norms – when the belief matched the action. However, there were also notable differences across the experiments, with more preference explanations produced overall when the belief did not match the action, more preference explanations produced when the norm was statistical, and inconsistent effects of norm typicality across different norm types. Again, these differences between Experiment 1 and Experiment 2 could be due to differences between producing explanations rather than rating provided explanations. In particular, participants may have been more inclined to provide preference explanations in Experiment 2 when norm or belief explanations were inadequate because the preference explanation was deemed a sufficient answer, and correspondingly been less inclined to note a preference when a belief or norm explanation was adequate, resulting in more dramatic effects for norm and belief agreement.

Alternative causal history explanations. The frequency of alternative causal history (ACH) explanations produced across conditions is presented in Figure 2.3. ACH explanations were more frequent when the norm matched the action, $c^2(1) = 41.067$, $p < .001$. However, this effect was qualified by a higher-order interaction with norm type, $c^{2(}2) = 15.85$, $p < .001$: Norm-action agreement did not have a discernable impact for moral norms.

ACH explanations were also more frequent when the belief mismatched the action, $c^2(1) = 6.34$, $p = .012$, potentially because explanations that appealed to belief in the norm were less viable.

In sum, ACH explanations were impacted both by whether norms and whether beliefs matched the action, occurring more frequently when the norm matched the action and also when the belief mismatched the action. This is consistent with the possibility that norms are believed to causally influence behavior via mediating mental states or processes, but that these need not involve a correct and explicit belief about the norm. When this explicit belief is absent, participants are reasonably happy to accept norm explanations (as seen in Experiment 1), but additionally posit or assume an alternative causal pathway between the norm and the action, as reflected in ACH explanations.

However, this explanation for the effect of norm-agreement does not apply equally to all norm types, as the influence of norms varied as a function of norm type: ACH explanations were more common for conventional and statistical norms when the norm matched the action than when it mismatched, but this was not the case for moral norms. This could be because participants in the relevant condition (N+B-) were quite willing to provide norm explanations in the moral case (see Fig. 2.3), consistent with hints from findings across both experiments that moral norm explanations are relatively "belief-independent," and accepted and provided in the absence of explicitly worked out mental states mediating the relationship between norms and actions.

Relationship between explanation production and other dependent measures. For each of the final four dependent measures, we compared average ratings across participants who provided a given explanation type with those who did not (see Table 2.4). We highlight two notable results. First, participants who provided norm or belief explanations provided higher ratings than those who did not for the strength of the agent's preference for the outcome of his action. This pattern does not correspond to that

from Experiment 1, but reinforces the suggestion that the explanation production task may have encouraged participants to produce a single sufficient explanation even when they accepted a role for alternative factors, since norm explanations were associated with a greater role for preferences.

Second, participants' who provided ACH explanations rated the agent's preference significantly lower than those who did not. This is consistent with the idea that ACH explanations are produced when more canonical mental state explanations fail – such as those involving a belief in the norm or a desire for the outcome. Surprisingly, however, the same relationship was not observed for ratings of the strength of the agent's belief in the norm.

Summary of key results. The results from Experiment 2 reveal that norm explanations are spontaneously produced when the norm matches the action being explained. For the case or moral norms, in particular, such explanations were offered even when the agent whose actions were being explained did not have a corresponding belief in the norm. These findings mirror those from Experiment 1, but Experiment 2 provides additional insight into why norm explanations may have been produced under such conditions: the production of explanations containing alternative causal histories relating the norms to behavior. For example, some participants appealed to unconscious beliefs or imitation. These causal histories were "alternative" in the sense that that did not require the agent's knowledge of the corresponding norm, but nonetheless recognized a causal role for the agent and his mental states in mediating the relationship between norms and behavior.

General Discussion

The results of these two studies suggest that norm explanations are judged suitable explanations for behavior, and that norm explanations are not always simply shorthand for a causal process that includes a corresponding belief in the norm. In particular, norm explanations were more likely to be endorsed and produced overall when accompanied by a corresponding belief, but were also frequently endorsed and produced in the absence of such a belief, especially when the relevant norm was moral. The free responses from Experiment 2 suggested that some participants generated alternative causal pathways through which the norm could influence behavior, bypassing the need for canonical mental states such as knowledge of the norm. In addition, both experiments found that norms influenced belief and preference explanations. In Experiment 1, belief explanations were rated more highly when accompanied by a norm that matched behavior, while preference explanations were rated more highly when a behavior was norm violating. Experiment 2 involved a more complex relationship between norms and belief explanations, but similarly found that preference explanations were more frequent for norm-violating actions. And finally, both experiments found differences in norm explanations as a function of norm type. Below we consider the implications of these findings and further elaborate on the differences found across different types of norms.

What is the Relationship Between Norms and Behavior?

Our results demonstrate that people both accept and provide norms as explanations for behavior when the norm matches the behavior, and sometimes even do so in the absence of an accurate belief about the norm – at least in cases involving moral norms. These results challenge the causal model depicted in Figure 2.1, according to

which the influence of norms on an agent's action is mediated by corresponding beliefs and desires. If this causal process doesn't underlie norm explanations for behavior, what does?

One possibility is that participants recognize a role for mental states or other mediating factors between a norm and behavior, but don't believe that the specific mental state we manipulated – an accurate belief that the norm is the case – is necessary. We call this the "enriched model" (see Figure 2.4a). This model is consistent with the alternative causal history explanations generated in Experiment 2, which cited alternative causal pathways involving unconscious beliefs, cultural influences, or imitation, among others. However, the idea that people reason with a relatively complex and sophisticated causal model of the relationship between mental states and behavior is at odds with the findings cited in the introduction (Apperly et. al, 2006; Lin et al., 2010), which instead suggest that mental state inferences are often bypassed or performed retrospectively as needed.

An alternative that's more consistent with prior work is depicted in Figure 2.4b, and identified as the "sparse explanatory schema" model. The basic idea is that people typically operate with a sparse causal model that involves an implicit commitment to mediating mental states, but relies heavily on norms. This model can be employed without actually working out the specific mental states involved, instead filling them in only if and when needed. In addition to the work of Apperly et. al (2006) and Lin et al. (2010) already cited, this alternative is supported by a phenomenon known as the "illusion of explanatory depth" (Rozenblit & Keil, 2002). Adults not only possess relatively sparse causal models (Keil, 2003), but seem to find them explanatorily adequate, only appreciating the sparseness of their model when called on to actually explain (Rozenblit & Keil, 2002). It could be that people reason with the sparse explanatory schema depicted in Figure 2.4b without appreciating that they are doing so, reverting to the enriched model (Figure 2.4a) only when called upon to explain, as in Experiment 2.

Although the spare explanatory model is quite skeletal, it can nonetheless be motivated from the fact that norms often serve as reliable bases for predicting and explaining behavior. Norms provide reasons for action (Searle, 2001), so in the absence of evidence for a competing belief or desire, a norm provides good evidence for corresponding beliefs and desires that in turn guide behavior. In fact, 5-year-olds (but not older children or adults) will actually favor norms over desires in predicting an agent's behavior (Kalish & Shiverick, 2004). With adults, Malle and colleagues (O'Laughlin & Malle, 1998; Malle, 1999) have found that explanations for behavior often invoke reasons (a species of mental state), but also what they call "causal histories of reasons." The latter explanations cite factors that are not typically themselves mental states, but rather bring about reasons for actions, such as situational clues, childhood experiences, and, importantly for our purposes, cultural training and norms (see also Uttich, Tsai, & Lombrozo, in preparation). Causal history of reasons explanations are especially common when explaining the behavior of groups, which can involve heterogeneous or unknown mental states, making the norm (or other causal contributor to reasons) the common factor with a reliable relationship to behavior. For the same reasons that explainers may favor causal history of reason explanations, norm explanations that don't specify mental states may be attractive.

If the sparse explanatory schema model is correct, then our results don't indicate that participants actively reject the basic model (Figure 2.1) or actively endorse the enriched model (see Figure 2.4a). Instead, participants typically employ a sparse model (Figure 2.4b) that is perfectly adequate for many everyday judgments, but can specify mental states and alternative causal pathways when required.

Alternative Accounts

One alternative possibility is that people represent norms as having a direct causal influence on behavior. In this type of model, norms act like natural laws, such as gravity, which can influence an agent whether or not the agent *believes* the law is true of the world. This possibility seems unlikely as a general account of norms, however – by age 5, children appropriately differentiate the role of knowledge and desires in conforming to natural laws versus social rules (Kalish, 1998).

A final possibility is that norms are accepted as direct explanations for behavior not as a result of people's causal commitments concerning the relationship between norms and behavior, but instead as the result of a performance problem in applying this model. This possibility appeals to the competence/performance distinction in linguistics introduced by Chomsky (1965). On this account, people endorse a causal model that includes beliefs as a mediating factor between norms and behavior (Figure 2.1 or 2.4a), but additional factors lead people to ignore or misapply this model. Similar arguments have been made to account for children's Theory of Mind development and performance on the false belief task (Wimmer & Perner, 1983), where one perspective is that the concepts needed to understand false belief are present from an early age, but that young children fail because of other demands imposed by the test (Fodor, 1992.) In our task, it could be that participants failed to correctly apply the model in Figure 2.1 (or that in Figure 2.4a), and instead resorted to a simpler strategy, such as picking out as the explanation whatever information matched the agent's behavior. This possibility seems unlikely, however. Our task was very straightforward, did not involve time pressure, and required participants to make a single judgment, so it isn't clear what the performance problem could be. Moreover, the data aren't easily explained by a simple alternative, such as accepting or providing any "matching" information, since (among other things) this fails to account for systematic responses concerning preference explanations, and also systematic differences across norm types, to which we now turn.

Differences Between Norm Types

Previous research has distinguished between prescriptive and descriptive norms as well as between moral and conventional norms (Smetana, 1981; Turiel, 1983). Our findings partially track these distinctions, as different kinds of norms generated somewhat different patterns of results.

First, several effects of norm type corresponded to differences between conventional and descriptive (statistical) norms. Norm explanations were given the highest ratings when the norm was conventional, and the lowest ratings when the norm was statistical. The opposite pattern, with low ratings for conventional norms and high ratings for statistical norms, was observed for preference explanations in Experiment 1. And in Experiment 2, preference explanations also occurred most often when the norm was statistical. These results make sense in light of differences across prescriptive and descriptive norms, where only prescriptive norms provide "desire-independent" reasons for action (Searle, 2001). So independently of an agent's desire, she has a reason to

conform to a prescriptive norm. In contrast, statistical norms describe behavior rather than providing a reason for action (or, potentially, provide a weaker reason for action if conforming to statistical norms is itself a motivating desire). As a result, prescriptive norm explanations can more safely bypass mental states where statistical norm explanations require an inference and appeal to a supporting desire.

Our findings also revealed some ways in which moral norms may be unique. One of the more robust findings concerning moral norm explanations was an interaction between norm type and belief-action agreement. For participants in Experiment 1 presented with moral norms, belief explanation ratings were insensitive to whether the belief matched the action. Similarly, in Experiment 2, participants presented with moral norms were equally likely to generate norm explanations when the belief didn't match the action as they were when the belief did match. These patterns were not observed for conventional or statistical norms. Moral norms also generated a unique pattern for preference explanations: Participants in Experiment 1 presented with moral norms rated preference explanations more highly when beliefs matched the agent's action, with the opposite pattern for conventional and statistical norms. And in Experiment 2, participants presented with moral norms were more likely to produce preference explanations when the belief and the action of the agent matched than when it did not, a pattern that was not observed for the other two types of norm. Why did moral norms differ in these ways?

As alluded to in the introduction, previous research suggests that many people regard moral norms as universal (Turiel, 1983) and objective (Goodwin & Darley, 2008). Moral objectivism is roughly the idea that moral claims are true in a way that does not depend on the beliefs, values, feelings or practices of others (Harman, 1975; Sinnott-Armstrong, 2009). While attempts to empirically investigate people's beliefs concerning moral objectivism have yielded somewhat variable results, there is support for the idea that many people are moral objectivists about at least some moral norms (Nichols, 2004; Goodwin & Darley, 2008; Sarkissian et al., 2011; see Knobe et al., 2012, for review). If participants in our studies tended to view the moral norms as true facts about the world that did not need intervening beliefs to hold true, they may have been more inclined to cite them as explanatory factors in the absence of particular beliefs. Consistent with this idea, variation in objectivist beliefs appears to be related to a particular kind of norm explanation: Participants who are more objectivist about a given moral norm are more likely to accept an explanation for historical change that appeals to that norm – for example, explaining that slavery ended "because slavery is morally wrong" (Uttich, et al., in preparation).

Nonetheless, some aspects of our results challenge an appeal to moral objectivism as the exclusive explanation for the unique properties of moral norms in explanations. Moral objectivism would arguably predict an effect of norm typicality, since participants would presumably only extend objectivist assumptions to typical norms that they themselves recognize as true. Norm typicality did interact with norm type, but only in the production of preference explanations in Experiment 2 – not where an appeal to moral objectivism would make an interaction seem most likely, more heavily penalizing the atypical norm explanations in the moral case as compared to the conventional and statistical cases. Nor it is obvious why moral objectivism would yield the observed effects of norm type on preference explanations across Experiments 1 and 2. Moral objectivism can potentially explain some of the unique effects of moral norms in our

studies, but not all of them.

Conclusion

We investigated norm explanations as a window onto people's causal models of the relationship between norms, mental states, and behavior. Our findings suggest that norm explanations are frequently endorsed and provided when behavior matches the norm, and – at least for moral cases – even when the relationship between the norm and behavior is not mediated by a corresponding belief. We've suggested that people often employ a sparse explanatory schema involving unspecified mental states, as this schema captures the reliable relationship between norms and behavior in many everyday cases, and can therefore support effective explanations and predictions. However, people do seem to appreciate the mediating role of mental states, and can elaborate the sparse schema when necessary. This approach to Theory of Mind reasoning captures people's remarkable sophistication when it comes to reasoning about behavior, but also helps explain our cognitive efficiency and occasional shortcomings when it comes to mental state inferences in everyday contexts.

Figures

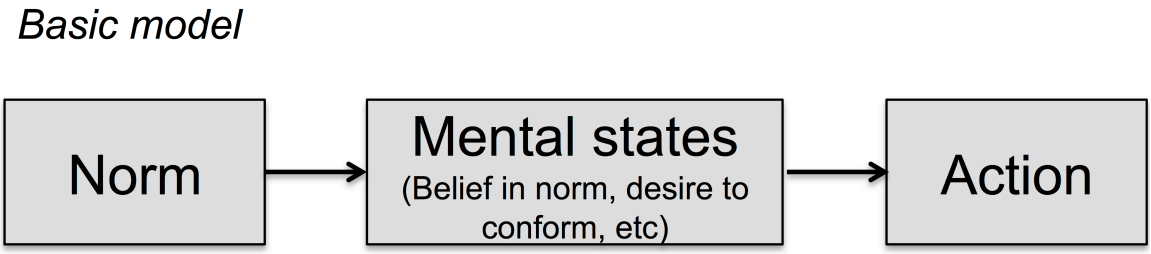


Figure 2.1 - Candidate causal model for the relationship between norms and behavior, where norms can causally impact actions, but only when mediated by appropriate mental states.

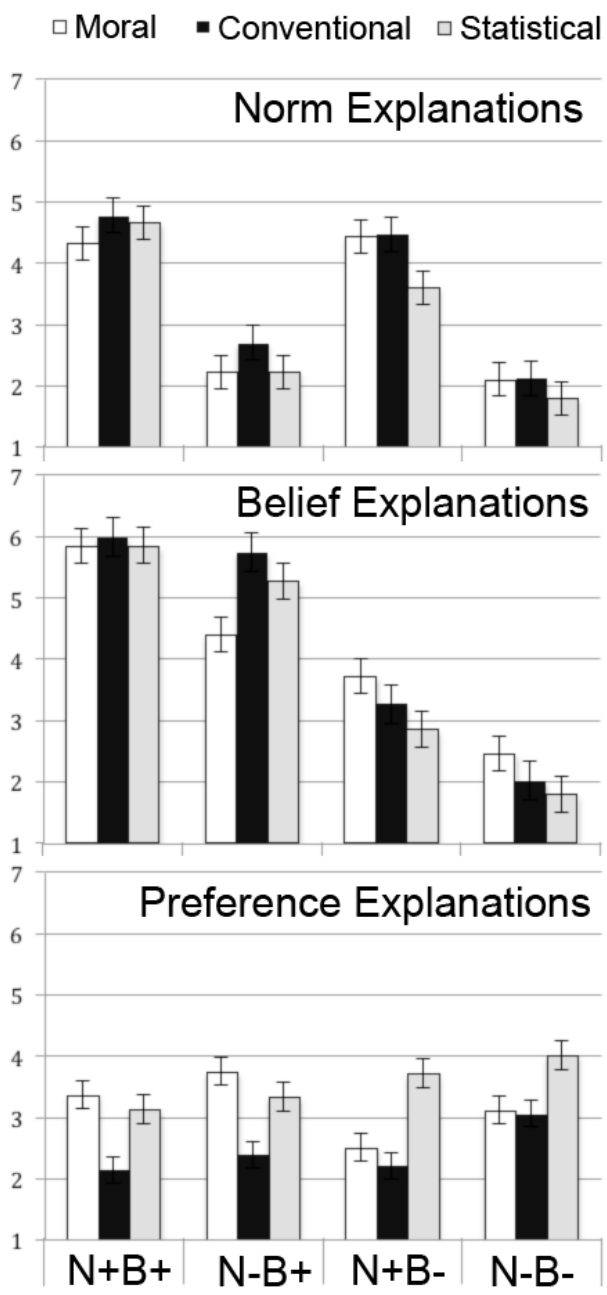


Figure 2.2 - Average norm explanation ratings, belief explanation ratings, and preference explanation ratings for Experiment 1 as a function of norm-action agreement ("N+" indicates a match, "N-" a mismatch), belief-action agreement ("B+" indicates a match, "B-" a mismatch), and norm type. Error bars correspond to one standard error of the mean.

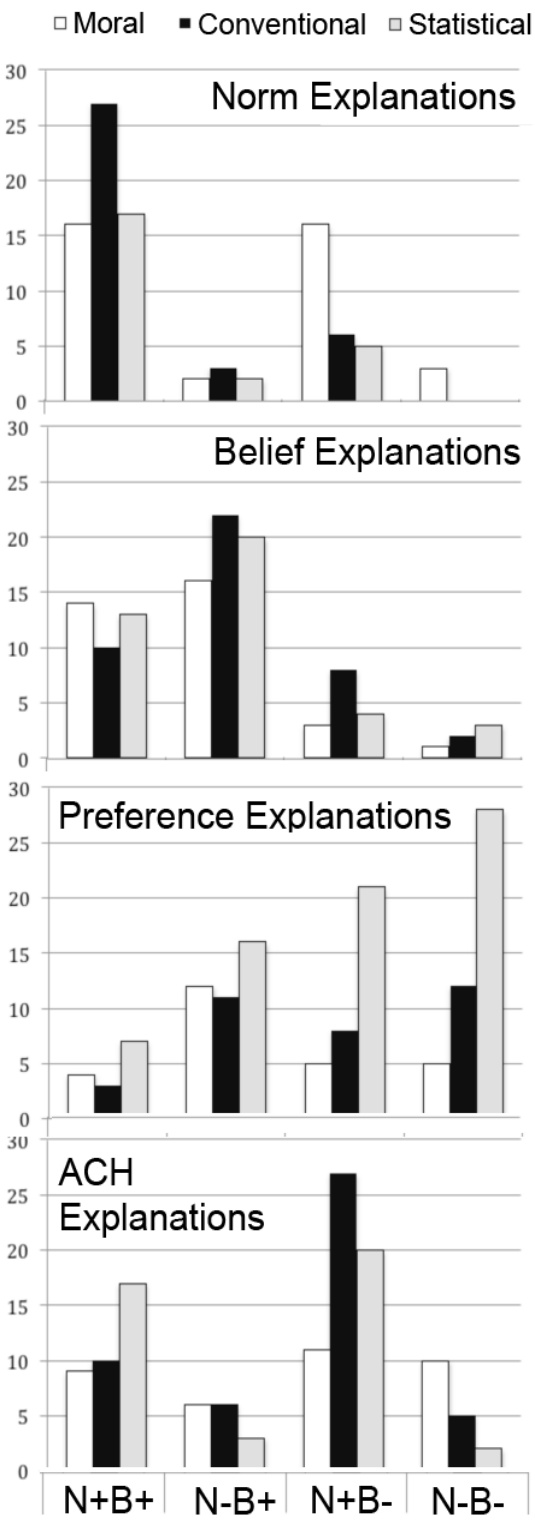


Figure 2.3 - Total norm explanation counts, belief explanation counts, preference explanation counts, and alternative causal history (ACH) explanation counts for Experiment 2 as a function of norm-action agreement ("N+" indicates a match, "N-" a mismatch), belief-action agreement ("B+" indicates a match, "B-" a mismatch), and norm type.

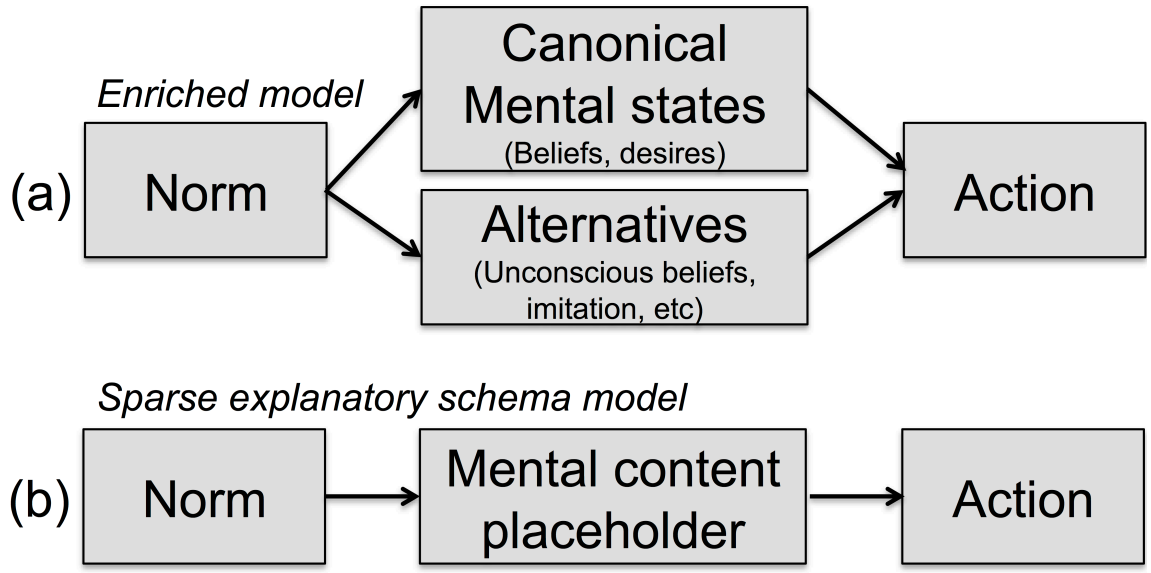


Figure 2.4 - Graphical depiction of two possible causal models that could account for our results.

Tables

Moral	Disagreement resolution norm	Punishment norm
Typical	Respond to conflict with diplomacy before violence.	Discipline child by verbal scolding.
Atypical	Respond to conflict with violence before diplomacy.	Discipline child by spanking.
Conventional	Fork placement norm	Funeral dress norm
Typical	Place to the left of the plate.	Wear black.
Atypical	Place to the right of the plate.	Wear white
Statistical	Shoe color norm	T-Shirt color norm
Typical	Buy black shoes.	Buy light color t-shirts.
Atypical	Buy orange shoes.	Buy dark color t-shirts.

Table 2.1: List of norm types and topics for Experiments 1 & 2. Both typical and atypical norms are included.

Fork placement norm	Norm Match Belief Match (N+B+)	Norm Mismatch Belief Match (N-B+)	Norm Match Belief Mismatch (N+B-)	Norm Mismatch Belief Mismatch (N-B-)
Typical	Norm – Left Belief – Left Action - Left	Norm – Left Belief – Right Action - Right	Norm – Left Belief – Right Action - Left	Norm – Left Belief – Left Action - Right
Atypical	Norm – Right Belief – Right Action - Right	Norm – Right Belief – Left Action - Left	Norm – Right Belief – Left Action - Right	Norm – Right Belief – Right Action - Left

Table 2.2: List of possible norm, belief, and action combinations for the fork placement norm used in Experiments 1 & 2. Both typical and atypical norms are included.

	Norm	Believe Norm	Preference	Alternative Causal History
Moral	"Because it is the norm of his culture."	"Because he believe that it is the norm to respond so."	"He is comfortable with this discipline."	"He learned it was the way to deal with the problem at hand."
Conventional	"Bob placed the fork to the left of the plate because he was raised in a culture in which it is the norm to set the fork to the left of the plate. "	"Bob believed that placing the fork on the right of the plate was the "Toma" cultural norm."	"Maybe he wanted to do something different for a change."	"Peer pressure."
Statistical	"because it is a social norm for the Toma culture."	"He thinks that orange shoes are the norm."	"Bob just likes light t-shirts."	"Looking like the rest of his group reinforces Bob's status as a member."

Table 2.3 - Examples of explanations provided in Experiment 2 from each coding category. Participants' original spelling and grammar are preserved.

	Provided **norm** explanation?		Provided **belief** explanation?		Provided **preference** explanation?		Provided **ACH** explanation?	
Strength of…	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**
agent's preference	4.90**	4.31	4.84**	4.30	4.59	4.35	3.98**	4.52
agent's belief about what norm should be	4.58**	3.89	3.66*	4.09	3.67**	4.10	4.21	3.94
your belief about what norm should be	3.91	3.55	3.64	3.60	3.15**	3.75	3.55	3.63
your belief about what norm is	4.18	3.79	3.76	3.88	3.82	3.87	3.88	3.85

Table 2.4- Mean ratings for measures concerning the agent's and the participant's beliefs from Experiment 2 as a function of whether the participant provided a particular type of explanation or not. The means were analyzed with t-tests, and significant differences are indicated with asterisks (*<0.05, **<0.01).

Chapter 4: Ethical Explanations, Moral Objectivity, and Moral Progress [7]

In this chapter we investigate whether ethical or moral norms can serve as explanations for social-historical events. In addition we examine whether or not beliefs about the nature of morality impacts the use of these ethical explanations. In examining these questions we attempt to address whether an individuals understanding and reasoning about a moral norm impacts the relationship of the norm to their social-cognitive capacity for explaining past events.

Introduction

Why was slavery abolished in the United States? One type of explanation could cite economic or political factors – for example, the fact that the industrial revolution made slavery economically unfeasible, or that the North won the Civil War. Another type of explanation could cite historical changes in individuals' or demographic groups' broader beliefs about the morality of slavery, thus resulting in a rejection of the practice. But such social-historical explanations or appeals to changing beliefs do not exhaust the possibilities. Consider, for example the following quote from emancipated slave William Williams: "I believe the North will whip the South, because I believe they are in the right" (Blassingame, 1977, also cited in Cohen, 1997). This explanation for why the North would win the Civil War implicitly cites a moral norm ("they are in the right") and as such is an instance of what we will refer to as an ethical explanation.

Ethical explanations cite moral norms to account for particular social facts or changes. Importantly, the norms are offered as *explanations* for particular social facts or changes, not as *evaluations* of those facts or changes. Other examples of ethical explanations include explaining the extension of voting rights to women by stating that denying women the right to vote is morally wrong, or explaining the (potential, future) legalization of same-sex marriage by noting that restricting marriage to heterosexual couples is unjust. More mundane, smaller-scale examples could include explaining the return of a found wallet by claiming that "it was the right thing to do," or the decision to donate to charity by suggesting that it is a moral obligation to help those in greater need.

In this paper, we examine whether and why people accept ethical explanations. Most psychological and philosophical accounts of explanation suggest that explanations cite one or more causes (e.g., Salmon 1984; Woodward, 2003), identify a function or goal (e.g., Lombrozo & Carey, 2006; Wright, 1976), provide underlying mechanisms (e.g., Bechtel, 2008; Craver, 2007; Darden, 2006; Ahn & Kalish, 2000), and/or appeal to descriptive generalizations, such as physical laws (see Woodward, 2010 for review from philosophy; see Lombrozo, in press for review from psychology). But at first glance, ethical explanations do not seem to fit these accounts. In particular, existing accounts of explanation focus on *descriptive* information, while ethical explanations involve an inherently normative component. Ethical explanations appeal to what *ought* to be, rather than any particular description of the state, mechanism, or causal force relevant to what is being explained.

[7] The material in chapter 2 was co-authored with Dr. Tania Lombrozo and Dr. George Tsai.

We hypothesize that (some) people conceptualize moral norms in such a way that they do satisfy the requirements for an adequate explanation. For example, conceptualizing moral norms as directed forces that have a causal impact on individual or group behavior would render them somewhat analogous to natural laws. Just as one can explain a an increase in entropy by appeal to thermodynamic laws, one might explain the demise of slavery by appeal to what is "right," where some principle of justice is regarded as a kind of directed force that leads societies from a less moral to a more moral state. In other words, some people could have meta-ethical commitments – commitments about the nature of morality – that support the legitimacy of ethical explanations. The meta-ethical commitments that we examine here are moral objectivism – roughly the belief that moral norms exist independently of particular individuals' or groups' beliefs – and moral progress – the idea that social-historical change tends towards moral improvement. In particular, we test the hypothesis that people who endorse moral objectivity and moral progress are more likely to endorse and provide ethical explanations.

Ethical explanations

Within philosophy, the legitimacy of ethical explanations is controversial. For example, Joshua Cohen argues that ethical explanations can legitimately account for social and historical change, and in particular that the injustice of American slavery itself contributed to its eventual demise (Cohen, 1997). David Brink argues that the "political instability and social protest in [apartheid] South Africa" is better explained in terms of racial oppression (an unjust practice) rather than in terms of the particular social, economic, and political conditions that were instantiated in South Africa (Brink, 1989). But many philosophers, historians, and social scientists question whether ethical explanations can be adequate explanations of social and historical facts (Leiter, 2001; Harman, 1977; Williams, 1985). What grounds this negative view is roughly the thought that moral facts—e.g., facts about the injustice of practices such as slavery or the restriction of voting rights to men—could not themselves contribute to explaining the demise of these systems. In particular, many deny that moral facts could be causal factors in social and historical change.

One possibility is that what appear to be ethical explanations – such as those concerning slavery offered above – are in actuality shorthand for more typical explanations that cites people's *beliefs,* where the beliefs in question concern what is ethical. So an explanation such as, "slavery ended because it is wrong" might in fact be shorthand for an explanation citing changes in people's beliefs as the cause or mechanism of slavery's demise – e.g., "slavery ended because people came to believe it is wrong." On this view, once we look below the surface grammar of people's explanations, these explanations are not in fact ethical explanations, but what we term "ethical belief" explanations.

A second possibility is that ethical explanations are conceptualized as teleological or functional explanations that specify a goal or end state. On this view, an explanation such as "slavery ended because it is immoral" would be roughly equivalent to "slavery ended in order to make the world a more moral place." Note, however, that this possibility does not make the acceptance of ethical explanations any less mysterious, as it remains unclear how an ethically-defined endpoint could explain social or historical change.

A more intriguing possibility is that the surface grammar of people's explanations of social facts reveals a set of meta-ethical commitments according to which moral norms are in fact responsible for social-historical change. The two classes of meta-ethical commitments that we consider are moral objectivism and moral progress. If, in fact, people endorse ethical (and not only ethical belief) explanations, they could be more inclined to do so to the extent they also endorse moral objectivity and moral progress.

Moral objectivity

For the purposes of our investigation, to accept that morality or moral norms are objective is to believe that some moral claims are true in a way that does *not* depend on people's decisions, feelings, beliefs, or practices. To reject the objectivity of moral norms—i.e., to accept moral skepticism of one kind or another—is therefore to deny that there are any moral claims that are true or to allow that moral claims are true in a way that depends on decisions, feelings, beliefs, or practices (Harman, 1975; Sinnott-Armstrong, 2009). Cultural or moral relativism is an example of the latter position. For a moral relativist, ethical *belief* explanations could be especially appropriate, as they identify the basis for moral norms and are also causes of human behavior. For a moral objectivist, the moral norms cited in ethical explanations could be conceptualized more like natural laws, and therefore be regarded as appropriate explanations for social-historical change.

To illustrate the relationship between objectivism and the role of norms in explanations, consider possible explanations for historical changes in aesthetic versus scientific domains. For the visual arts, one might explain why paintings have become less figurative over time by noting that they are consequently more expressive. To the extent that people are relativists about aesthetic judgments (Goodwin & Darley, 2008; Kuhn, 2001; Wainryb et al., 2004), however, they could regard the presupposition that abstraction is more expressive as a mere, subjective assumption and instead favor an explanation that cites another historical change, such as the development of photography. In contrast, one might explain a biological change, such as why a species' genetic variation decreases in small populations over time, by appeal to a natural law or process, such as genetic drift. To the extent that people are objectivists about science, this explanation is likely to be satisfactory and not subject to the concerns that arise for visual arts.

Recently, there have been a few attempts to examine empirically what people believe about moral objectivity (Nichols, 2004; Goodwin & Darley, 2008; Sarkissian et al., 2011; see Knobe et al., in press, for review). For example, Goodwin and Darley (2008) asked participants to rate their agreement with factual, ethical, social-conventional, and taste statements, and then asked them whether these statements were true, false, or an opinion or attitude. For example, one of the ethical statements was "Robbing a bank in order to pay for an expensive holiday is a morally bad action," while "Wearing pajamas and bath robe to a seminar meeting is wrong behavior" was a social-conventional statement. Participants were later asked whether the fact that someone disagreed with them about a given statement meant that the other person was wrong, that neither person was wrong, that they themselves were wrong, or something else entirely. In a second study participants were asked to provide a "yes" or "no" answer to the question, "According to you, can there be a correct answer as to whether this statement is true?" Using these measures Goodwin and Darley found evidence that people treat

statements of ethical beliefs as more objective than either social conventions or taste. They also found a great deal of variation in objectivism across both ethical statements and individuals. Strongly held ethical beliefs were seen as more objective than beliefs people did not hold as strongly, and those who said they grounded their ethical beliefs in religion, moral self-identity and/or the pragmatic consequences of failing to observe norms were more likely to be objectivist about ethical statements.

More recently, Sarkissian et al. (2011) have argued that relativistic beliefs are more prevalent than suggested by Goodwin and colleagues, but that these beliefs are only observed when participants are comparing judgments made by agents who are very different from each other in important ways. In their studies, participants were presented with two agents who disagreed about a moral claim and were asked whether one of them must be wrong. For example, participants were asked to imagine a race of extraterrestrial beings called Pentars who "have a very different sort of psychology from human beings," and were then asked if a classmate and a Pentar had differing views on a moral case to rate their agreement with the statement "at least one of them must be wrong." Participants provided more objectivist answers ("one of them must be wrong") when comparing judgments made by agents from the same culture, but more relativistic answers (denying that "at least one of them must be wrong") when comparing judgments made by agents from different planets (a human and a Pentar). Sarkissian et al. argue that engagement with radically different perspectives is the psychological process that led participants to moral relativism.

In sum, recent research points to the interest and viability of studying meta-ethical commitments concerning moral objectivism, but the different results generated for different moral norms and with different measures leaves many questions unanswered. For our purposes, previous research on moral objectivism provides a set of useful measures of moral objectivity and demonstrates variation in the extent to which people endorse moral objectivity. Such findings are relevant background for assessing the hypothesis that variation in the acceptance of ethical explanations is related to how strongly an individual endorses moral objectivity.

Moral Progress

The second meta-ethical commitment we consider concerns the idea that history tends toward moral progress — that is, that there is a general tendency for the morally relevant features of the world to improve over time. This notion, which postulates a certain directionality in human history, can be contrasted with the notion of mere moral change. Although there have been philosophers in the history of philosophy, notably Marx and Hegel, who have defended the idea that human history tends toward moral progress, the notion also finds expression in people's ordinary or common-sense thinking. For example, Martin Luther King famously proclaimed, "the arc of the moral universe is long but it bends towards justice" (King, 1986).

It is worth noting that a belief in a historical tendency towards moral progress can be consistently held while maintaining that moral progress can be imperceptible and can occur over long stretches of time. Sometimes moral improvement can be dramatic and rapid, while at other times not. Thus, belief in a tendency towards moral progress does not require commitment to a particular rate of moral progress. Additionally, to hold that there is a basic tendency towards moral progress in human history is also compatible with allowing that these tendencies do not inevitably or necessarily prevail. Believing in some

tendency need not require belief in inevitability. For example, one could believe that six-year-old children tend to grow physically larger (e.g., that a child at 14 years of age will be larger than that very same child at age 6) without claiming that they inevitably or necessarily get physically larger (after all, serious illness or death could prevent their continuing to grow in size). Likewise, in the case of moral progress, one could still allow that there could be exogenous forces such as environmental and biological catastrophes or foreign invasions that prevent the historical development towards moral progress.

While previous research has not (to our knowledge) investigated beliefs about moral progress directly, there is a large body of research on a related notion referred to as "belief in a just world" (Lerner, 1980; Furnham, 2003). Belief in a just world refers to the idea that good things happen to good people while bad things happen to bad people. The belief that people experience consequences that correspond to the moral nature of their actions or character is potentially consistent with belief in moral progress, although the relationship is complex. For example, it is not obvious that the world is morally improved when a criminal develops cancer, unless the prevention of future criminal activity (through deterrence, incapacity, or death) is the moral payoff. We return to the relationship between belief in a just world, moral progress, and ethical explanations in our second experiment.

Overview of experiments

In the following studies, we first demonstrate that ethical explanations are indeed frequently endorsed and provided in accounting for social-historical change. In particular, we concentrate on cases of explaining historical changes concerning social, political, and legal institutions and practices – namely slavery, women's suffrage, and same-sex marriage. We then turn our attention to trying to make sense of why people frequently endorse ethical explanations. To this end, we examine whether there is a relationship between participants' acceptance (or rejection) of ethical explanations, their acceptance (or rejection) of the idea that morality is objective, and their acceptance (or rejection) of the notion that human history tends toward moral progress. In Experiment 2, we additionally examine the relationship between the generation of ethical explanations and belief in a just world.

In addition to allowing us to explore the status of ethical explanations and their relationship to meta-ethical beliefs, our investigation of ethical explanations serves two additional aims. First, examining the relationship between the acceptance of ethical explanations, people's beliefs about moral objectivity, and people's beliefs about moral progress provides an opportunity to explore the range and variation in moral beliefs across individuals, as well as the internal consistency of different moral beliefs within individuals. Second, the status of ethical explanations is a source of contemporary debate within philosophy that has received little empirical attention. We return to these additional aims in the general discussion.

Experiment 1

In order to determine whether ethical explanations are accepted as legitimate explanations for social-historical facts, we first examine whether people will endorse them as good explanations when they are provided. In particular, we examine whether ethical explanations are rated as being better than explanations that are designed to be

non-explanatory, and on par with explanations that cite non-ethical causes or ethical beliefs, which are generally recognized as legitimate kinds of explanations.

If ethical explanations are accepted, we can then attempt to answer *why* some might consider these explanations appropriate by examining the relationship between the acceptance of ethical explanations and participants' meta-ethical views. We hypothesize that participants who endorse moral objectivism and moral progress are more likely to accept ethical explanations than those who do not endorse these meta-ethical views.

Participants

Ninety-six participants (57 female, mean age 33) were recruited from Amazon Mechanical Turk, an online crowd-sourcing platform. Participants received a small payment for their participation. All participants identified themselves as being from the United States and as native speakers of English. There were 12 conditions, with 8 participants in each condition.

Materials and procedure

The experiment consisted of five main parts: (1) explanation ratings, (2) moral objectivity measures, (3) moral progress measures, (4) self-reported meta-ethical commitments, and (5) baseline check on morality of social fact. The ordering of the parts was partially counterbalanced, as detailed below after the measures are introduced.

Explanation ratings. Participants read a question that asked why some historical change occurred or will occur, such as "Why was slavery abolished?" Participants were presented with one question from among three cases of historical (or potential future) change in the U.S.: (1) the abolition of slavery (which has already occurred), (2) women's suffrage (which has already occurred), or (3) the legalization of same-sex marriage (which has not occurred in the U.S. as a whole as of this writing).

After reading the question, participants rated four possible explanations for the change. The explanations were of four different types. *Ethical explanations* cited moral norms. *Ethical belief explanations* cited changes in people's beliefs or attitudes about the morality of the social phenomenon in question. *Non-ethical explanations* cited some economic or social event. Finally, *Poor explanations* were circular or trivial explanations.[8] The order in which participants rated each explanation was randomized. Sample explanations for the slavery prompt are included below (explanation type labels were not presented to participants):

> Why was slavery abolished in the United States?
>
> Ethical: Because slavery is morally wrong.
> Ethical belief: Because people came to believe owning slaves was morally wrong.
> Non-ethical: Because of the emergence of the industrial revolution in the United States.
> Poor: Because owning slaves was made illegal.

Participants rated the explanations on a 1-7 scale with 1 being a "poor explanation," 7 a "good explanation," and 4 an "average explanation." For the same-sex

[8] In a pilot phase of the experiment, participants on Amazon Mechanical Turk rated candidate explanations for the *ethical*, *ethical belief*, and *non-ethical* categories. The top-rated explanation from each category for each social fact was used in Experiment 1.

marriage condition, participants were asked about a potential future event rather than a past event, namely "Why might same sex marriage be allowed in the United States?"

Moral objectivity measures. Participants' views concerning moral objectivity were examined in two different ways. The first involved an adaptation of the method used by Sarkissian et. al. (2011). Participants read vignettes where either a person similar to themselves (i.e., from their same time and culture) or a person from another time period (e.g., the 18th century) disagreed with an imagined friend of the participant about whether a social fact was morally problematic. The relevant social fact was always matched with that for which participants had rated explanations. Examples from the slavery condition are presented below:

> Imagine a person named Allison, a fairly ordinary student from your town who enjoys watching sports and hanging out with friends. Consider Allison's views concerning the moral status of the following social institution: Slavery.
>
> Allison thinks that slavery is not morally wrong.
>
> Imagine that one of your friends thinks that slavery is morally wrong. Given that these individuals (Allison and your friend) have different judgments about this case, we would like to know whether you think at least one of them must be wrong, or whether you think both of them could actually be correct. In other words, to what extent would you agree or disagree with the following statement concerning such a case?
>
> "Since your friend and Allison have different judgments about this case, at least one of them must be wrong."

And:

> Imagine the social world of the United States in the 18th century. Most people in this time and place view slavery as morally acceptable. The existence of slavery is seen by many as part of the natural social order, slavery is permitted by the law and the slave trade is at its peak, and someone who owns many slaves is esteemed as admirable.
>
> An individual, Jessica, from this society (18th century United States), regards slavery as not morally wrong.
>
> Imagine, also, that an individual who is one of your friends thinks that slavery is morally wrong. Given that your friend and Jessica have different judgments about this case, we would like to know whether you think at least one of them must be wrong, or whether you think both of them could actually be correct. In other words, to what extent would you agree or disagree with the following statement concerning such a case?
>
> "Since your friend and Jessica have different judgments about this case, at least one of them must be wrong."

Participants rated their agreement with these statements on a 1-7 scale with 1 being "definitely disagree," 7 "definitely agree," and 4 "neither agree nor disagree." The order of the vignettes, current or historical first, was counterbalanced across participants.

Participants' beliefs about moral objectivity were also examined using a method adapted from Goodwin and Darley (2008). Participants were asked whether statements about the moral permissibility of the social facts are true, false, or an opinion. The example from the slavery case is below:

> The statement "slavery is not morally wrong" is:
>
> 1. True
> 2. False
> 3. An opinion

Participants answered questions concerning moral beliefs for all three historical facts, with the historical fact for which they rated explanations presented first.

Moral progress measures. Participants rated their agreement with statements concerning the tendency towards moral progress and whether moral progress is inevitable. Three statements examined participants' belief about the moral direction of human social change with questions concerning the abstract idea of moral progress. Additionally, participants were asked three questions about the inevitability of all three of the specific social historical changes. The order in which participants rated statements was randomized. All six questions about moral progress are listed below:

> Human history tends to move in the direction of a more perfect moral world.
> As time goes on, humanity does NOT generally become more morally advanced.[9]
> Over time we will move towards a more morally just world.
> The demise of slavery was inevitable.
> The extension of the right to vote to women was inevitable.
> The legalization of same sex marriage is inevitable.

Participants rated the statements on a 1-7 scale with 1 being "definitely disagree," 7 "definitely agree," and 4 "neither agree nor disagree."

Self-reported meta-ethical commitments. Finally, participants also completed two "self-report" questions concerning their meta-ethical commitments to objectivism or relativism. Unlike the previous measures of moral objectivity, these questions attempted to assess more directly participants' explicit beliefs. The first question, which we will refer to as the relativism scale, asked:

> How much do you agree with the following statement: Moral claims depend on decisions, feelings, beliefs or practices of a group or culture.

Participants rated their agreement with this statement on a 1-7 scale with 1 being "definitely disagree," 7 "definitely agree," and 4 "neither agree nor disagree." The second

[9] This item was reverse-coded.

question asked them to choose whether they believed they were better described as moral objectivists or moral relativists:

> Please choose which of the following best describes your beliefs about moral questions:
>
> 1) I believe in moral objectivism: Moral claims are true or false and do not depend on things like cultural factors or beliefs.
> 2) I believe in moral relativism: The truth of a moral claim depends on cultural factors like what a particular group believes.
> 3) Neither of these describes my beliefs about moral claims.

Baseline check. Participants were also asked for their personal views on whether the three social changes (demise of slavery, women's suffrage, legalization of same-sex marriage) were good or bad. For example, for the slavery fact participants were presented with the following statement:

> The demise of slavery was a good thing.

Participants rated their agreement with this statement on a 1-7 scale with 1 being "definitely disagree," 7 "definitely agree," and 4 "neither agree nor disagree." All three social facts were rated. The social fact related to the explanation each participant had rated was always presented first.

Counterbalancing Participants either first rated explanations (part 1) and then completed the moral objectivity measures (part 2) and the moral progress measures (part 3), with the order of parts 2 and 3 counterbalanced, or they first completed the moral objectivity measures (part 2) and the moral progress measures (part 3), with order counterbalanced, followed by the explanations ratings (part 1). Participants always completed the self-reported meta-ethical commitment measures (part 4) and check on social facts (part 5) last.

Results

We begin by reporting the data for each set of questions individually, and then consider the relationship between different meta-ethical commitments and between explanation ratings and meta-ethical commitments.

Initial analyses

Explanation ratings. In order to examine whether participants endorsed ethical explanations for social-historical changes, explanation ratings for the four explanation types were compared. Paired samples t-tests comparing explanation types revealed that ethical explanation ratings were not significantly different from ethical belief explanations, $t(95) = .800$, $p = .426$, nor from non-ethical explanation, $t(95) = 1.246$, $p = .216$. Ethical explanations were, however, rated significantly more highly than poor explanations, $t(95) = 2.169$, $p < 0.05$ (see Figure 3.1). These results suggest that ethical explanations are indeed endorsed by participants at the same level as ethical belief and non-ethical explanations, which are generally recognized as legitimate kinds of explanations for social-historical change.

Moral objectivism measures. Participants' moral objectivism was measured with two scales. The first measure, the average rating for both the current and historical judgments adapted from Sarkissian et al. (2011), yielded an average rating of 4.70 out of 7 (SD = 1.69), with higher scores indicating a greater belief in objectivism. The average ratings for the individual items were 4.82 (SD = 1.795) for the current case and 4.58 (SD = 1.833) for the historical case. A repeated-measures ANOVA with time period (current versus historical) as a within-subjects factor and social fact (slavery, women's suffrage, same-sex marriage) as a between-subjects factor did not reveal and significant effects.

The second measure of moral objectivism, the true, false, or opinion measure, yielded 18 "true" responses (6%), 102 "false" responses (35%), and 162 "opinion" responses (56%) out of 288 total responses (96 participants x 3 social facts). The distribution of true/false versus opinion responses did not differ significantly across the social facts, $\chi^2(2) = 3.69$, p = .16.

Moral progress measures. A moral progress measure, consisting of the average rating for all six questions concerning tendency towards progress and the inevitability of particular changes, was calculated for each participant. The average rating was 4.53 out of 7 (SD = 1.11), with higher scores indicating a greater belief in moral progress. This moral progress measure was also broken down into two subscores. The "tendency towards progress" subscore ("tendency" subscore, for short) consisted of the average rating for the three questions concerning a tendency towards moral progress. The average score was 3.68 out of 7 (SD = 1.36), with higher scores indicating a greater belief in the tendency towards moral progress. The "inevitability" subscore consisted of the average score for the three items concerning the inevitability of particular social-historical changes, and was 5.38 out of 7 (SD = 1.47), with higher scores indicating a greater belief in the inevitability of social-historical change. A repeated-measures ANOVA revealed no significant differences across the three social facts considered. The correlation between tendency and inevitability subscores across participants was .215 ($p < 0.05$).

Self-reported meta-ethical commitment measures. The self-reported relativism scale yielded an average of 5.28 out of 7 (SD = 1.58), with lower scores indicating a greater belief in moral objectivism. When asked to select whether their views were best described as relativist, objectivist, or neither, 29 participants (30.2%) self-identified as objectivist, 36 (37.5%) self-identified as relativist, and 31 (32.3%) indicated "neither." Unsurprisingly, relativism scores differed significantly as a function of participants' self-identified label as determined by a one-way ANOVA, $F(2,93) = 10.272$, $p < 0.01$. A Tukey post-hoc test revealed that ratings for self-identified objectivists (M = 4.36, SD = 1.97) differed significantly from those for self-identified relativists (M = 6.03, SD = 0.910, $p < 0.01$). The average ratings for those who identified as "neither" (M = 5.26, SD = 1.37) did not differ from ratings for either self-identified objectivists ($p = 0.053$) or self-identified relativists ($p = 0.081$). Given the large number of responses in the third category ("neither"), however, the self-report measure is not used in subsequent analyses.

Relationships between meta-ethical commitment measures. The two measures of moral objectivity and the self-reported relativism scale ratings were significantly correlated, with the average current/historical score correlated with the true/false/opinion

measure (r = -.294, $p < 0.01$) and the relativism scale (r = -.334, $p < 0.01$).[10] The true/false/opinion measure and relativism scale were also correlated (r = .311, $p < 0.01$). These correlations are significant and in the predicted directions. The correlations are small enough to suggest that each measure captures some unique variation in belief, but large enough to make differentiating the relationship between each measure and ethical explanation ratings problematic. Because the current/historical judgment was always presented before the true/false/opinion measure and self-reported relativism, we use only this measure in the analyses that follow. The current/historical rating also provides a continuous measure that can be motivated on the basis of past research (Sarkissian et al., 2011) as a measure likely to capture both objectivist and relativist views.

Current/historical judgments were not correlated with moral progress score (r = 0.076, p = 0.46) or with either subscale (tendency: r = 0.041, p = 0.69, inevitability: r = 0.079, p = 0.44).

Baseline check measures. The baseline check indicated that the moral norms presented were overwhelmingly accepted by participants. The average ratings were 6.13 of 7 (SD = 1.571) for the demise of slavery, 6.61 (SD = 1.070) for women's suffrage, and 5.72 (SD = 2.106) for same sex marriage. Agreement with the social fact that matched the explanations rated by a given participant was 6.66 (SD = 0.98). These ratings verify that participants endorsed the examined social changes and are not considered further. We do note, however, that ethical explanation ratings were (not surprisingly) correlated with the agreement score for the corresponding social fact (r = .274, $p < 0.01$): participants were more likely to endorse a norm as an explanation if they endorsed the norm.

Explanation ratings & meta-ethical commitments.

To examine whether participants' meta-ethical beliefs contributed to their endorsement of ethical explanations, a regression analysis was performed on explanation ratings using the current/historical moral objectivity score, the moral progress score, and an interaction factor as predictors. Ethical explanations were significantly predicted by current/historical moral objectivity score, b = .228, $t(95)$ = 2.370, $p < 0.05$, and also by moral progress, b = .286, $t(95)$ = 2.976, $p < 0.01$ (R^2 = 0.144, $F(2,95)$ = 7.820, $p < 0.01$). In other words, participants' ratings of ethical explanations increased as their endorsement of moral objectivism and moral progress increased. There was no significant interaction, b = -.032, $t(95)$ = -.324, p= .747 (see Figure 3.2).

Given that moral progress was a significant predictor for ethical explanation ratings, we ran a subsequent regression to examine the contributions of each moral progress subscale, moral tendency versus moral inevitability. The regression indicated that the inevitability subscale significantly predicted ethical explanations, b = .344, $t(95)$ = 3.465, $p < 0.01$, but tendency towards progress did not, b = .039, $t(95)$ = 0.388, p = 0.699. That is, ratings of ethical explanations increased as participants' endorsement of the inevitability of social-historical change increased, but endorsing a tendency towards moral progress did not have a comparable effect.

[10] Current/historical correlations are negative due to reversal of the scales (an objectivist would score high on current/historical, and low on the true/false/opinion measure and the relativism scale).

Ratings for ethical *belief* explanations were also predicted by current/historical moral objectivity score, $b = .195$, $t(95) = 2.003$, $p < 0.05$, and by moral progress , $b = .273$, $t(95) = 2.794$, $p < 0.01$, with no interaction ($R^2 = 0.121$, $F(2,95) = 6.374$, $p < 0.01$): participants' ratings of ethical explanations increased as their endorsement of moral progress and moral objectivism increased. A regression analysis using only the two moral progress subscales, moral tendency and moral inevitability, again indicated that the inevitability subscale significantly predicted ethical belief explanation ratings, $b = .296$, $t(95) = 2.937$, $p < 0.01$, but tendency towards progress did not, $b = .068$, $t(95) = 0.673$, $p = 0.503$.

Non-ethical explanations were not predicted by either measure (current/historical moral objectivity score, $b = -.123$, $t(95) = -1.202$, $p = .232$; moral progress: $b = -.145$, $t(95) = -1.418$, $p < 0.16$; $R^2 = 0.033$, $F(2,95) = 1.606$, $p = .206$.

Finally, poor explanations were predicted by moral progress, $b = -.265$, $t(95) = 2.68$, $p < 0.01$, but not by current/historical moral objectivity score, $b = .193$, $t(95) = 1.955$, $p = 0.054$, with no significant interaction, $b = .049$, $t(95) = .491$, $p = .625$ ($R^2 = 0.099$, $F(2,95) = 5.134$, $p < 0.01$). Participants' ratings for poor explanations *decreased* as their belief in moral progress increased. A regression analysis using only the two moral progress subscales, tendency and inevitability, indicated that the tendency towards progress significantly predicted poor explanation ratings, $b = -.338$, $t(95) = 3.379$, $p < 0.01$, but that the inevitability subscale did not, $b = .009$, $t(95) = 0.088$, $p = 0.930$. In other words, when participants endorsed a tendency towards moral progress, poor explanation ratings were lower.

These results suggest that both of the meta-ethical beliefs we examined independently impact participants' ratings of ethical and ethical belief explanations, but not non-ethical explanations. Additionally, rejecting moral progress predicts higher ratings for poor explanations.

Discussion

Experiment 1 provides evidence that participants generally judge ethical explanations to be as good as other explanations for social-historical facts, and that participants' meta-ethical beliefs about moral objectivity and moral progress are related to ethical explanation ratings. More specifically, both greater objectivism (as reflected by the current/historical measure) and greater belief in the inevitability of moral progress predict higher ratings for ethical explanations.

Meta-ethical beliefs also predicted ratings for the ethical belief explanations and for poor explanations. At first blush it could appear worrisome that the relationship between meta-ethical beliefs and ethical belief explanations mirrors that for ethical explanations. However, the fact that meta-ethical scores predict ethical belief explanation ratings could further support the idea that ethical explanations are the result of a particular conception of the nature of morality. Ethical belief explanations were included to address the argument that ethical explanations should not be taken at face value and are really belief explanations. However, if it were simply the *belief content* and not the *ethical content* of an explanation that mattered, it is unclear why we would see meta-ethical measures as predictors for ethical belief explanations. That meta-ethical beliefs predict ethical belief explanation ratings therefore suggests that even when *belief* in a moral norm is included in an explanation, the ethical part of the explanation remains explanatory.

Additionally, the negative relationship between moral progress and poor explanation ratings suggests that for those who do not believe in moral progress, the poor explanations were not so poor. Perhaps this is because those who do not endorse moral progress do not find the poor explanations to be as circumstantial, as what made the poor explanations poor was a potentially circular reference to the act of making a practice illegal. Perhaps these circular explanations are capturing something more explanatory if one does not believe the moral world was already likely to move in that direction.

This study is the first (to our knowledge) to empirically investigate ethical explanations. However, there is previous work relevant to intuitive beliefs about both moral objectivism and (more indirectly) moral progress. While differences in methodology make direct comparisons to past work on moral objectivism difficult, we found significant variation in beliefs about the objective nature of morality, as did Goodwin and Darley (2008). Additionally, the small though significant correlations between our different measures of moral objectivity support a conclusion from Sarkissian et al. (2011), namely that different measures of moral objectivity can yield different judgments, either because each measure captures different facets of people's meta-ethical beliefs or because the measure itself influences how people conceptualize ethical norms. The lack of a significant difference between the current and historical measures of moral objectivity in our study could also indicate that the historical case is an intermediate of sorts between the "normal" and "alien" cases of Sarkissian et al. (2011). With respect to moral progress, previous research has considered related beliefs, such as belief in a just world. This measure is introduced and examined in Experiment 2.

Experiment 2

Experiment 1 establishes that people endorse ethical explanations when they are provided, but does not address whether people spontaneously *provide* ethical explanations when asked to explain social-historical events. If some participants do provide such explanations, we can then examine whether those who provide ethical explanations are more likely to have particular meta-ethical beliefs. As in Experiment 1, we hypothesize that those who provide ethical explanations are more likely to endorse moral objectivity and moral progress.

In addition to examining the explanations participants provide, Experiment 2 goes beyond Experiment 1 in taking a more fine-grained look at the relationship between belief in moral progress and ethical explanations. The two subscales for moral progress in Experiment 1, tendency and inevitability, were confounded with the abstractness or concreteness of the described change. All three items probing belief in a tendency towards moral progress were abstract, such as: "Over time we will move towards a more morally just world." In contrast, all three items about the inevitability of moral progress were about particular social-historical events, such as: "The demise of slavery was inevitable." It is thus unclear whether the relationship between the inevitability subscores and ethical explanation ratings, which drove the effect of moral progress for both ethical and ethical belief explanations, was the result of the corresponding items' emphasis on inevitability or the use of concrete examples. Inevitability and concreteness are deconfounded in Experiment 2 by including statements that describe a concrete tendency towards moral progress (e.g., "Over time there is moral progress concerning slavery") as

well as those that involve an abstract statement of inevitability (e.g. "An increase in moral justice in the world is inevitable").

Finally, Experiment 2 examines the specificity of the relationship between meta-ethical beliefs and ethical explanations by considering yet another kind of meta-ethical belief: belief in a just world (BJW). Previous research suggests that BJW, the belief that good things happen to good people and bad things happen to bad people, is a prevalent and stable commitment (Furnham, 2003). If strongly endorsing any commitment concerning an ethical structure to the world is sufficient to find ethical explanations explanatory, then high BJW should predict the production of ethical explanations. If, in contrast, the relationship between meta-ethical commitments and ethical explanations is more finely coordinated, BJW could be irrelevant to the kinds of explanations we consider. For other kinds of explanations invoking moral content (e.g., "she won the lottery because she deserved to," "he became ill because he is a bad person,"), BJW could indeed be relevant. This additional measure thus allows us to examine whether the endorsement of ethical explanations has a relatively diffuse relationship to a host of different meta-ethical beliefs, or is instead more narrowly associated with beliefs about objectivity and progress.

In sum, Experiment 2 was similar to the previous experiment, with the following changes. First, participants provided free response explanations rather than rating provided explanations. Second, participants evaluated a more extensive set of questions concerning moral progress, allowing us to differentiate judgments of a progressive tendency from inevitability. Third, participants completed a standard scale of belief in a just world (Dalbert, Montada, & Schmitt, 1987). Finally, we removed the self-reported meta-ethical commitment measures, which were not particularly fruitful in Experiment 1.

Participants

Three-hundred-eighty-four participants were recruited from Amazon Mechanical Turk. Participants received a small payment for their participation. All participants identified themselves as being from the United States and as native speakers of English. There were 24 conditions, with 16 participants in each condition.

Procedure

The experiment consisted of four main parts: (1) explanation ratings, (2) moral objectivity measures, (3) moral progress measures and general belief in a just world measure (GBJW), (4) baseline check on morality of social fact. The ordering of the parts was partially counterbalanced, as detailed after the measures are introduced.

Explanations of Social Facts. Participants read a question that asked about why some change occurred or will occur (slavery, women's suffrage, same-sex marriage), as in Experiment 1. Participants were prompted to answer the question "in a sentence or two" and typed explanations into a text box.

Moral objectivity measures. The moral objectivity questions adapted from Sarkissian et al. (2011) were the same as those in Experiment 1. Once again, participants only answered questions corresponding to the social fact for which they provided explanations. The order of the vignettes (current or historical first) was counterbalanced across participants.

The questions adapted from Goodwin and Darley (2008) were adjusted to a likert scale format. Participants rated their agreement with statements concerning the moral permissibility of a social practice (e.g., "slavery is not morally wrong") on a 1-7 scale

with 1 being "is best described as true or false," 7 "is best described as an opinion", and 4 "is equally well described as either true/false or as an opinion." Participants rated all three social facts for these types of questions.

Moral progress measures. Participants again rated their agreement with statements concerning the tendency towards moral progress and the inevitability of all three historical changes. However, participants were also presented with three new "abstract" statements about the inevitability of moral progress, and three new "concrete" questions about moral progress concerning the three social facts:

> The moral advancement of humanity is NOT inevitable.[11]
> It is inevitable that on average our morals will be better than those of our distant ancestors.
> An increase in moral justice in the world is inevitable.
> Over time there is moral progress concerning slavery.
> Over time there is moral progress concerning voting rights.
> Over time there is moral progress concerning marriage rights.

Just world measures. Participants evaluated six statements concerning belief in a just world, taken from the General Belief in a Just World Scale (GBJW; Dalbert, Montada, & Schmitt, 1987). These questions are included below.

> I think basically the world is a just place.
> I believe that, by and large, people get what they deserve.
> I am confident that justice always prevails over injustice.
> I am convinced that in the long run people will be compensated for injustices.
> I firmly believe that injustices in all areas of life (e.g., professional, family, politic) are the exception rather than the rule.
> I think people try to be fair when making important decisions.

Participants rated the statements on a 1-7 scale with 1 being "definitely disagree," 7 "definitely agree," and 4 "neither agree nor disagree." The order of all moral progress and GBJW questions was randomized.

Baseline check. As in Experiment 1, participants were asked for their personal views on whether the three social changes were good or bad.

Counterbalancing. Participants either provided explanations of social facts (part 1) and then completed the moral objectivity measures (part 2) and the moral progress and GBJW measures (part 3), with the order of parts 2 and 3 counterbalanced; or they first completed the moral objectivity measures (part 2) and the moral progress and GBJW measures (part 3), with order counterbalanced, followed by providing explanations of social facts (part 1). Participants always answered the check on social facts last (part 4).

Results

We begin by reporting the coded explanations participants provided and the data for each set of measures individually, and then consider the relationship between

[11] This item was reverse-coded.

different measures of meta-ethical belief and between explanation categories and meta-ethical commitments.

Initial analyses

Explanations provided. Two coders coded responses for whether they contained one or more of the four explanation types from Experiment 1: ethical, ethical belief, non-ethical, and poor. Any explanation involving an ethical or moral statement as an explanation was coded as ethical. An ethical belief explanation referenced a belief in some ethical or moral quality. Poor explanations were those explanations that repeated some form of the question, with no additional explanatory information. All other explanations were coded as non-ethical explanations. In infrequent cases, responses included multiple explanations of different types; these explanations were coded in multiple categories. However, an explanation that included a belief in a moral norm, without a direct appeal to the norm itself as an explanation, was only coded in the ethical belief category. Table 3.1 includes sample explanations from each category.

Agreement between the two independent raters was 91% (Kappa = .787, $p < .01$) for the ethical explanations. As a conservative strategy to avoid overestimating the prevalence of ethical explanations, an explanation was only coded as ethical if both coders agreed. For other explanations types, disagreements were resolved by discussion. Interrater reliability was 87% (Kappa = 0.681 ($p<.01$) for all explanations.

A quick look at the proportion of ethical explanations provided, 88 ethical (23%), 116 ethical belief (30%), 215 non-ethical (56%), and 21 poor (6%), suggests that participants do spontaneously provide ethical explanations when asked about social-historical changes (see Table 3.2). McNemar's test reveals that the percentage of ethical explanations provided was significantly lower than that for non-ethical explanations, $p < 0.01$, and higher than that for poor explanations, $p < 0.01$, but not significantly different from that for ethical belief explanations, $p = 0.056$. A comparison between explanations that contained any ethical component (ethical or ethical belief) and non-ethical explanations revealed no significant difference, $p = 0.507$

Moral objectivism measures. The average rating for the combined current/historical moral objectivism measure was 4.72 out of 7 (SD = 1.87), with higher scores indicating a belief in moral objectivism. The average rating for the current vignette alone was 4.84 (SD = 1.95) and that for the historical vignette alone was 4.62 (SD = 1.97). A repeated-measures ANOVA with time period (current versus historical) as a within-subjects measure and social fact (slavery, women's suffrage, same-sex marriage) as a between-subjects measure revealed two significant effects: a main effect of time period, $F(1,381) = 13.17$, $p < .01$, with more objectivist responses for the current vignette than for the historical vignette, and a main effect of social fact, $F(2,381) = 36.35$, $p < .01$, with responses that were more objectivist for slavery (M = 4.99, SD = 1.90) and women's suffrage (M = 4.90, SD = 1.69) than for same-sex marriage (M = 4.30, SD = 1.95).

The average rating for the true/false/opinion measure was 4.31 (SD = 2.15), with lower scores indicating greater moral objectivism. This measure also varied as a function of social fact, $F(2,382) = 53.65$, $p < .01$, with the most objectivist responses for slavery (M = 3.71, SD = 2.58), followed by women's suffrage (M = 4.31, SD = 2.51) and same-sex marriage (M = 4.91, SD = 2.30).

Moral progress measures. The average scores for each set of three moral progress items were as follows: 4.09 out of 7 (SD = 1.35) for abstract tendency towards

moral progress, 5.25 (SD =1.15) for concrete tendency towards moral progress, 3.89 (SD = 1.49) for abstract inevitability of social change, and 5.24 (SD = 1.38) for concrete inevitability of social change. The measures were also combined to allow for abstract vs. concrete and tendency vs. inevitability comparisons. Paired samples t-tests indicate that both factors generated significant differences: abstract items generated significantly lower ratings than concrete items (M=3.98, SD=1.32, versus M=5.24, SD=1.11, $t(383) = -21.666$, $p<.01$), and tendency was endorsed more strongly than inevitability (M=4.67, SD=1.10 versus M=4.56, SD=1.23, $t(386)= 2.385$, $p<.05$). We refer to the combined measure from all four subscales as moral progress-inevitability, or MPI. For correlations between the subscales see Table 3.3.

Belief in a just world measure. The average score for the six belief in a just world items was 4.0 (SD = 1.12), with higher scores indicating greater belief in a just world.

Baseline check measures. The baseline check indicated that the moral norms presented were overwhelmingly accepted by participants, with ratings of 6.70 of 7 (SD = .95) for the demise of slavery agreement, 6.63 (SD = 1.00) for women's suffrage, 5.15 (SD = 2.20) for same-sex marriage, and 6.06 (SD = 1.72) for the social fact that matched the explanations a given participant rated. A paired sample t-test revealed a significant difference between the matched agreement scores for those who provided ethical explanations (M = 6.51, SD = 1.01) versus those who did not (M = 5.94, SD = 1.87 $t(386) = -3.717$, $p < 0.01$)

Relationship between measures of meta-ethical commitments. Looking at the correlations between the three measures of moral objectivism we find average current/historical score correlated with the matching true/false/opinion rating ($r = .321$, $p < 0.01$) and with the average of all three true/false/opinion ratings ($r = .271$, $p < 0.01$). As in Experiment 1, we use the combined current/historical rating as an index of moral objectivity in subsequent analyses.

We also examined the correlations between the three measures of different meta-ethical commitments. The average current/historical score was correlated with the combined moral progress-inevitability score (MPI), $r = .120$, $p < 0.05$. MPI was also correlated with the general belief in a just world measure, $r = .433$, $p<0.01$. For a complete overview of correlations, including correlations between subscores, see Table 3.3.

Explanations provided & meta-ethical commitments.

Participants who provided ethical explanations differed significantly from those who did not on several meta-ethical measures (see Table 3.4). In particular, t-tests revealed significant differences in average current/historical ratings, with participants who produced ethical explanations providing ratings more consistent with a belief in objectivism. Additionally, the overall measure of moral progress (MPI), as well as the ratings for the four subscores individually, differed significantly between participants who did and did not provide ethical explanations, with those who provided ethical explanations providing higher ratings, consistent with a belief in moral progress. Belief in a just world did not differ significantly between participants who did and did not provide ethical explanations.

Additional t-tests revealed that participants who provided non-ethical explanations differed significantly from those who did not on a number of measures (see

Table 3.4). Differences were found for the overall measure of moral progress (MPI), as well as the ratings for the four subscores individually, and for the measure of belief in a just world. Those who provided non-ethical explanations provided lower ratings to both moral progress and belief in a just world, indicating that they endorsed these beliefs less strongly than those who provided other kinds of explanations.

Participants who provided ethical belief explanations did not differ significantly from those who did not, and those who provided poor explanations did not differ significantly from those who did not.

Discussion

The results from Experiment 2 suggest that participants do offer ethical explanations, and that those who offer ethical explanations have meta-ethical beliefs that differ from those who do not. In particular, participants who provided ethical explanations endorsed objectivism more strongly than those who did not (as assessed by the current/historical measure), and endorsed moral progress more strongly than those who did not (as assessed by the total moral progress-inevitability measure, as well as each individual subscale). However, General Belief in a Just World did not vary significantly between participants who did and did not provide ethical explanations. This suggests that it is the particular meta-ethical commitments originally investigated, belief in moral progress and moral objectivity, not just any meta-ethical commitments, that relate to ethical explanations.

Experiment 2 also found that participants who provided non-ethical explanations differed significantly from those who did not on measures of moral progress and General Belief in a Just World (GBJW). The difference for moral progress is potentially a consequence of the relationship between providing ethical explanations and endorsing moral progress: because participants generally provided only a single explanation type, those who provided ethical explanations very rarely provided non-ethical explanations as well. However, it is noteworthy that low GBJW was significantly associated with providing non-ethical explanations, while high GBJW was not significantly associated with providing ethical explanations.

It is also noteworthy that while Experiment 1 found similar relationships between meta-ethical commitments and both ethical and ethical belief explanations, Experiment 2 only found significant relationships for ethical explanations. It could be that participants differentiate ethical and ethical belief explanations less successfully when they are simultaneously provided (Experiment 1) than when they are asked to generate explanations themselves (Experiment 2). Alternatively, it could be that Experiment 2 was less sensitive than Experiment 1 and simply failed to find a relationship between meta-ethical commitments and ethical belief explanations.

Finally, the moral objectivity data from Experiment 2 support conclusions from both Goodwin and Darley (2008) and Sarkissian et al. (2011). Like Goodwin and Darley (2008), we found that judgments reflected greater objectivism for some social facts (slavery) than for others (same-sex marriage). Goodwin and Darley found that moral statements that were more strongly endorsed were more likely to yield objectivist judgments, and our data from the baseline check are consistent with this relationship, as the statement concerning slavery was endorsed more strongly than that concerning same-sex marriage. Like Sarkissian et al. (2011), we found that the judgments concerning whether two people with different moral views can both be right depend on the similarity

of the two people involved. Specifically, participants were more likely to provide objectivist responses for the current vignette than the historical vignette.

Returning to our central aims, the results from Experiments 1 and 2 successfully demonstrate that participants both endorse and spontaneously produce ethical explanations and that the endorsement and production of ethical explanations is linked to participants' meta-ethical beliefs concerning moral objectivity and moral progress.

General Discussion

The results of the two studies we present suggest that ethical explanations are indeed a part of people's reasoning about social-historical change. Participants in Experiment 1 endorsed ethical explanations on par with ethical belief and non-ethical explanations, and approximately 20% of the explanations provided by participants in Experiment 2 were ethical explanations.

The fact that ethical explanations are deemed genuinely explanatory is a bit puzzling for a few reasons. First, ethical explanations do not seem to meet the criteria for a good explanation, regardless of which of the competing criteria one considers. Ethical explanations seem to lack anything like a causal force, mechanism, or underlying, descriptive generalization. The second and related puzzle is why people would provide a normative answer to a question about a historical change, which on the face of it seems to call for a descriptive answer. We suggest that both of these puzzling issues have a common answer suggested by our results. Specifically, we believe our results suggest that because of their meta-ethical beliefs, some participants conceptualize moral norms in a way that makes them more analogous to generally accepted forms of explanation.

One possibility is that participants who endorse moral objectivism and/or moral progress regard moral norms as directed, causal forces that impact individual or group behavior. On this conceptualization of moral norms, ethical explanations can be regarded as causal explanations or as explanations that subsume a particular social-historical change under a law-like regularity (see Lombrozo, 2011, for review). Another possibility is that participants who endorse moral objectivism and/or moral progress regard ethical explanations in more functional or teleological terms – that is, as specifying an ideal, moral endpoint or goal – and believe that this goal satisfies the requirements for an adequate functional or teleological explanation (see Lombrozo & Carey, 2006 for a discussion of the psychological commitments that underlie functional/teleological explanations). Of course, these possibilities are not mutually exclusive and could apply differentially to moral objectivism versus moral progress. *Prima facie*, moral objectivism bears a closer relationship to the view that moral norms are akin to causes or natural laws, while moral progress involves an additional commitment to change in the direction of some specified ideal, consistent with a teleological explanation.

Even if some participants regard ethical explanations as perfectly legitimate, it is nonetheless puzzling that ethical explanations were so often preferred and provided over ethical belief explanations, which were always available and are arguably less problematic. One way to understand this preference is in terms of a distinction that Malle and colleagues draw between "reason" explanations and "causal history of reason" explanations (Malle, 1999). Reason explanations for why someone chose to work a 14-hour day, for example, could include that "he did so to make more money" or "because she wants to get ahead," while causal history of reason explanations would instead cite

antecedents to specific reasons, such as "he is driven to achieve" or "she works in a high pressure corporate environment." O'Laughlin and Malle (2002) found that causal history of reason explanations were provided more often for group behavior than for individual behavior, potentially because groups can have heterogeneous reasons. Our vignettes all involved group behavior in the form of social-historical changes, and could thus have pushed participants towards causal history of reason explanations over reason explanations that cited beliefs directly. In other words, for participants with particular meta-ethical views, ethical explanations could serve as good explanations because the moral norm identifies the causal background for a complex and heterogeneous set of reasons that underlie group behavior.

Our findings are also consistent with a growing body of work suggesting that everyday reasoning blurs the boundary between the normative and the descriptive. Recent research demonstrates that a variety of seemingly descriptive judgments, such as those concerning the cause of an event (Hitchcock and Knobe, 2009) or whether an action was carried out intentionally (Knobe, 2003a; Uttich and Lombrozo, 2010), are influenced by normative considerations, such as the moral status of an action's cause or outcome. For example, Hitchcock and Knobe (2009) presented participants with situations in which two factors were both necessary to bring about an outcome, but where one was consistent with a norm and the other violated a norm. For example, in one vignette, an intern and a doctor both decide to administer a drug to a patient, but the doctor notices that the drug can have severe side effects and signs off on it despite a hospital policy against administering the drug. Nonetheless, the patient's health improves. Hitchcock and Knobe found that participants overwhelmingly cited the norm-violating factor, the attending doctor rather than the intern, as the stronger or principal cause, revealing an influence of norms on causal attributions.

While there are competing theories about why normative considerations (such as moral status) impact seemingly descriptive judgments (such as causal and intentional ascriptions), current theories nonetheless concur in recognizing that beliefs about what *ought* to be the case are tightly tied to the process of reasoning in what are thought of as descriptive, non-moral judgments (such as causal reasoning and theory of mind). Perhaps ethical explanations are another instance of this broader phenomenon whereby normative considerations impact descriptive judgments, in this case explanations of social-historical change. However, in the case of ethical explanations, normative considerations not only influence or inform a descriptive judgment, such as which agent is identified as the cause of an event, but actually take the place of descriptive information, serving as explanations.

The basis for meta-ethical beliefs and other open questions

While we believe our results provide some insight into why normative information is provided by some participants when descriptive information is seemingly called for, there are still open questions concerning this process. For example, why do people hold the meta-ethical beliefs that they do? Is there a rational justification for ethical explanations given certain meta-ethical beliefs, or are people making an error? While we cannot answer these questions at present, our findings and those from previous research provide some guidance in moving towards answers.

Our results provide evidence that meta-ethical beliefs are diverse and complex. The results of our moral objectivism measures echo the conclusions from some of the

findings of the previous work done by Goodwin and Darley (2008) and Sarkissian et al. (2011). Like Goodwin and Darley (2008), we find significant variation in objectivism across individuals. Like Sarkissian et al. (2011), we find evidence that measures that highlight different perspectives seem to increase non-objectivist responses. Our findings support the need to consider the characteristics of both participants and measures in drawing conclusions about meta-ethical beliefs.

In addition to variation between individuals, there are open questions about the coherence and variability of meta-ethical beliefs *within* individuals. Correlations between our measures of meta-ethical beliefs suggest two conclusions: that the meta-ethical concepts we investigate have some psychological reality, but also that there is only partial consistency and coherence in these meta-ethical beliefs. The three separate measures of moral objectivity used in Experiment 1 and the two measures from Experiment 2 were all significantly correlated in the predicted directions, but only moderately so. Similarly, the moral progress subscores used in Experiment 2 as well as the belief in a just world measures were significantly correlated, but the correlations varied considerably in magnitude. If participants have strong, stable, and consistent meta-ethical commitments, why would responses to meta-ethical questions vary in the ways observed?

While our results do not directly address this question, they suggest three possible responses: the first largely methodological, the second concerning the coherence of meta-ethical commitments themselves, and the third the coherence of people's intuitive moral theories. The first response is in keeping with a suggestion by Sarkissian et al. (2011), who argue that when forced to address radically different perspectives, people who *appeared* to have completely objectivist commitments reveal some underlying, relativist intuitions. It could be that only some questions engage particular psychological processes and accurately assess the range of participants' commitments, with others obscuring their true range of beliefs. The second view argues that the meta-ethical commitments we investigate (a belief in moral objectivism or a belief in moral progress) do not correspond to coherent and unified sets of beliefs. Thus participants judgments could be somewhat inconsistent across measures because the philosophical constructs we aim to assess are themselves diverse or incoherent.

The third view suggests that participants' meta-ethical beliefs are best described as intuitive theories that are generally coherent, but need not be fully consistent (see, e.g., Thagard, 1989 on explanatory coherence). In work investigating the relationship between deontological versus consequentialist commitments as revealed in explicit versus indirect measures, for example, Lombrozo (2009) found evidence of a systematic but imperfect correspondence across measures. Perhaps responses to measures of meta-ethical beliefs concerning moral objectivity and belief in moral progress are similarly supported by semi-coherent intuitive theories. Like "folk physics" or intuitive theories of the supernatural, such theories will tend to support coherent judgments across tasks and measures, but can also generate inconsistent judgments given that intuitive theories are rarely explicitly articulated and examined (Shtulman, 2010). An attractive feature of this option is that it helps explain the relationship between meta-ethical commitments and ethical explanations: Supporting explanations is generally recognized as a core function of intuitive theories, so it makes sense that intuitive moral theories that embody meta-ethical commitments would be consulted in generating explanations for morally-relevant social change.

Further research aimed directly at measuring the nature of meta-ethical commitments can revisit these possibilities and address concerns about the stability of participants' meta-ethical commitments across both time and domain.

Philosophical Implications

Our findings do not show that ethical explanations are true or legitimate, but they do show that some people find ethical explanations to be explanatory. A natural question this raises is how the explanatory significance of (putative) moral norms should be understood, and how this is related to philosophical claims concerning the status of ethical explanations.

Philosophers have generally assumed that ethical explanations are inadequate—i.e., that they are not really explanatory. However, they have not had much to say in the way of reasons why ethical explanations are inadequate explanations. The fact that with a few exceptions, contemporary philosophers have simply (and wrongly) assumed that the ordinary folk do not seriously endorse ethical explanations could partially explain why they themselves have tended not to take ethical explanations seriously. Having shown that many people do in fact endorse ethical explanations, our hope is that this will motivate philosophers to say more about why ethical explanations are (or are not) inadequate. As Joshua Cohen has asked, "do ethical explanations withstand reflective examination, or are they simply collages of empirical rumination and reified hope, held together by rhetorical flourish?" (Cohen, 1997).

Given the fact that some people do accept and provide ethical explanations, and that this is correlated with their meta-ethical commitments, this raises philosophical questions about the adequacy of these explanations as well as their precise conceptual/logical relationship (if any) to the various meta-ethical commitments in question. For example, what is it exactly about the assumption that moral norms are *objective* that can lead a person who holds that assumption to suppose that moral norms could explain social change? If ethical explanations are deficient, is this to be explained wholly in terms of the meta-ethical commitments on which such explanations are often based? Are ethical explanations more defensible on the assumption that moral norms are objective?

There are also related philosophical questions about what ethical explanations imply about the relationship between the nature of morality/justice, on the one hand, and the nature of the social world, on the other. For example, are there sociological and historical laws to the effect that practices and social arrangements that fail to meet the norms of morality and justice will tend not to be viable, perhaps in the long term? What sort of evidence would count in favor of thinking such laws exist?

Finally, philosophers, going back to at least David Hume, have assumed that there is a rigid distinction between (historical) fact and moral value—in particular, between *historical description* and *explanation*, on the one hand, and *moral evaluation*, on the other. While there is, of course, a fundamental distinction between how the world is and how it would be good for the world to be, our study suggests that—at least for a segment of the population— people's explanations of history and their moral evaluations of history are not so insulated from each other.

Conclusion

Our findings reveal that people both endorse and provide ethical explanations, with those committed to moral objectivity or moral progress most likely to do so. The relationship between participants' meta-ethical beliefs and ethical explanations suggests that although ethical explanations appear at first glance to violate most psychological and philosophical accounts of explanation, certain meta-ethical beliefs can support a conceptualization of moral norms that allows them to serve as satisfactory explanations. The findings therefore bear on theories of explanation as well as questions in moral psychology concerning the variability in moral beliefs across individuals as well as the consistent of different moral commitments and judgments within individuals.

Figures

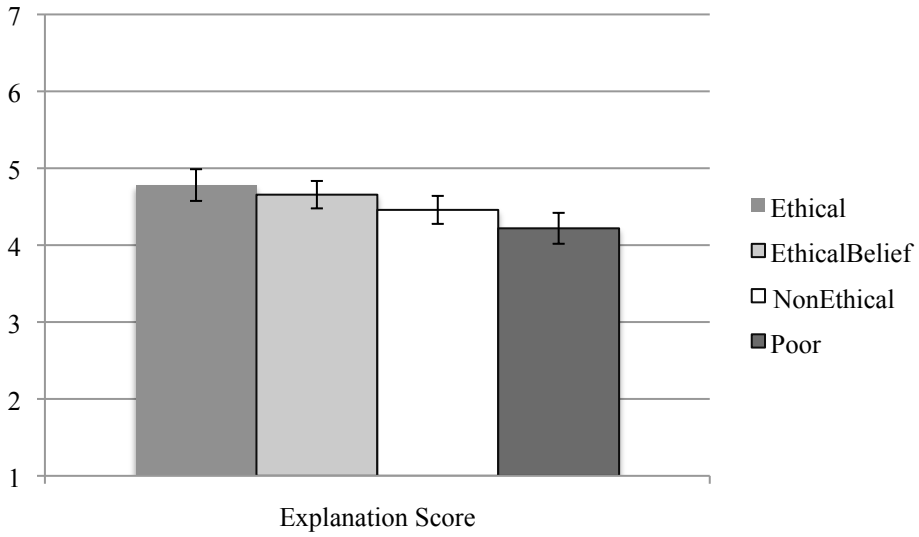


Figure 3.1- Explanation ratings from Experiment 1 on a scale from 1 ("poor explanation") to 7 ("good explanation"). Error bars indicate one standard error of the mean.

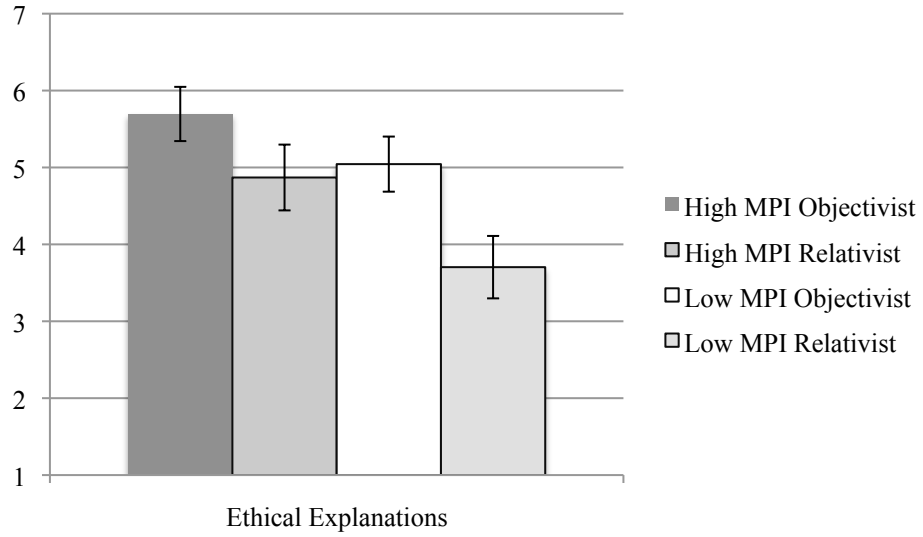


Figure 3.2- To visually illustrate the relationship between moral objectivity, moral progress, and ethical explanation ratings in Experiment 1, participants were classified into four groups based on their objectivity and progress scores. Those with scores above the median for the current/historical scale (4.5) were classified as objectivists and those below the median as relativists; Those above the median on the combined moral progress inevitability score (MPI, 4.5) were classified as "High MPI" and those below the median as "Low MPI." Average ethical explanation ratings are indicated for each group, with error bars corresponding to standard errors of the means.

Tables

	Ethical	Ethical Belief	Non-Ethical	Poor
Slavery	"Slavery was abolished by the United States because it is morally wrong."	"Salvery was abolished in the United States because it was morally wrong and enough people finally recognized it as such."	"Because it was no longer economically viable for the majority of Americans."	"By the enactment of the 13th Amendment to the Constitution"
Women's Suffrage	"because it was unfair to only have men vote and that both men and women should be treated the same."	"Because of the sufferage movement that eventually convinced society that it was a woman's right to vote."	"Because woman united in the sufferage movement and demanded an equal write to vote. They held marches and rally's until they were given this right."	"Because the Nineteenth Amendment was passed."
Same-Sex Marriage	"Perhaps because it's the morally correct policy?"	"Attitudes towards same sex marriage have shifted to become more positive especially among younger people."	"Because it is a faction of the voting public that can be very vocal."	"Because I can't think of a reason why not"

Table 3.1- Examples of explanations provided in Experiment 2 from each coding category. Participants' original spelling and grammar are preserved.

Explanation Type	All	Slavery	Women's Suffrage	Same-Sex Marriage
Ethical	88	34	23	31
Ethical-belief	116	47	27	42
Non-Ethical	215	67	79	69
Poor	21	10	6	5
TOTAL	440	158	135	147

Table 3.2- Coded explanations from Experiment 2. Cells correspond to the total number of each explanation type across all vignettes (first column) as well as subdivided for each individual vignette. Note that a single participants' explanation could fall into multiple coding categories, although this was rare.

	Current (C)	Historical (H)	C+H	Abstract Tendency	Concrete Tendency	Abstract Inevitability	Concrete Inevitability	MPI	Belief in a Just World
Objectivism Measures									
TF / opinion	.313**	.299**	.321**	-.078	.009	-.052	.060	-.021	.103*
Current (C)	1	.817**	.953**	.073	.225**	.102*	.052	.135**	.040
Historical (H)	---	1	.954**	.076	.157**	.085	-.003	.094	.071
C+H	---	---	1	.078	.200**	.098	.026	.120*	.058
Moral Progress Measures									
Abstract Tendency	---	---	---	1	.539**	.722**	.370**	.826**	.368**
Concrete Tendency	---	---	---	---	1	.510**	.522**	.780**	.281**
Abstract Inevitability	---	---	---	---	---	1	.465**	.858**	.455**
Concrete Inevitability	---	---	---	---	---	---	1	.737**	.266**
MPI	---	---	---	---	---	---	---	1	.433**

*<0.05, **<0.01

Table 3.3- Correlations between meta-ethical measures from Experiment 2.

	Provided **ethical** explanation?		Provided **non-ethical** explanation?	
	Yes	**No**	**Yes**	**No**
Objectivism Measures				
Current	5.26*	4.71	4.73	4.97
Historical	4.99*	4.51	4.45	4.83
AvgCH	5.13*	4.61	4.59	4.90
TF / opinion (matching)	4.27	4.68	4.62	4.54
TF / opinion (average)	4.16	4.35	4.33	4.28
Moral Progress Measures				
Abstract Tendency	4.38*	4.01	3.93**	4.29
Concrete Tendency	5.49*	5.17	5.12*	5.40
Abstract Inevitability	4.28**	3.77	3.64**	4.19
Concrete Inevitability	5.57**	5.14	5.10*	5.41
MPI	4.93**	4.52	4.45**	4.82
Additional Measures				
Belief in a Just World	4.16	3.96	3.84**	4.21
Moral Fact Agreement	6.50**	5.93	6.00	6.14

Table 3.4- Mean ratings for each measure from Experiment 2 as a function of whether participants provided an ethical explanation or not and whether they provided a non-ethical explanation or not. The means were analyzed with t-tests, and significant differences are indicated with asterisks (*<0.05, **<0.01).

Chapter 5: Conclusion

The preceding chapters have examined and attempted to characterize the relationship between social cognition and moral understanding. The impact of norm violation versus the influence of moral considerations was investigated for intentional action judgments in chapter two. Chapter three investigated the use of norms as explanations for the behavior of an agent as well as the impact of norm and belief information on other explanations of behavior. Building on this, chapter four examined the use of explanations that cite a moral belief to explain social-historical change and how the use of these explanations was related to beliefs about the moral world.

The three studies in chapter two used the side-effect effect to investigate whether and why conventional and moral norms influenced Theory of Mind judgments, namely ascriptions of intentional action. These studies demonstrated while there is an impact of moral status on judgments of intentional action the impact is not exclusive to moral norms. Norm violation influences judgments of intention concerning side effects and this holds when moral status is controlled, and when the norm is conventional. Additionally, norm-violating behavior has a greater impact on predictions of an agent's future behavior than does norm-conforming behavior. These findings together argue that there is a bidirectional relationship between moral understanding and Theory of Mind, but that it is best described as an evidential relationship. These findings also support the idea that this relationship is not just an issue of moral influence, but of norms informing mental state ascriptions.

Chapter three investigated the role of norms in explaining behavior by looking at participants' endorsement and production of explanations for events in vignettes that systematically varied norms, an agent's belief about norms, and actions of an agent. The results of the two studies in chapter three show that participants find norms to be suitable explanations for behavior and that their use as explanations is not just shorthand for a causal process that includes belief. Norms were both rated as acceptable explanations and provided as explanations by participants. The studies also suggested that the relationship between a norm and an action influenced the use of other potential explanations for the action. Additionally, the results indicate that while all norms seemed to have some impact, not all norms had the same effect. The results suggest there may be differences between prescriptive and descriptive norms, as well as differences within the prescriptive norm group between moral and conventional norms as they relate to explanations.

Chapter four considered whether and why people explain socio-historical change by appeal to moral norms, such as "slavery ended because it was morally wrong." The results of the two studies presented in chapter four suggest that explanations that cite a moral claim are a part of people's reasoning about social-historical change. These ethical explanations are both endorsed and provided at an equivalent rate to other explanations. Additionally, the studies suggest participants' meta-ethical beliefs, specifically participants' belief in either moral progress or in moral objectivism, is related to the variation in the use of ethical explanations. These results suggest that the way the moral world is conceptualized by participants impacts the way that they integrate moral norm information into their explanations for social-historical change.

Taken together the results of these studies argue that the relationship between social cognition and moral understanding is indeed bi-directional. Varying norm information impacted whether or not a side effect was judged to be intentional and affected explanations for behavior. Additionally, the variation seen in an individual's beliefs about the moral world was related to the way they explained social-historical events. This influence of normative reasoning, an evaluative process typically seen as part of moral understanding and not social cognition, on social-cognitive capacities suggests that the best way to characterize the relationship between moral understanding and social cognition is as a bi-directional relationship. In addition to the influence the information from social cognitive judgments can have on our moral reasoning, our normative understanding can provide a source of information useful in making social-cognitive judgments. Given this bidirectional relationship, we now turn to the two sub-questions raised in the introduction. First, given the influence of normative considerations on social-cognitive judgments, does this require a drastic revision of our understanding of social cognition or can we account for this normative influence while largely preserving our existing models of social cognition? Next, does the influence of norm information differ depending what type of norm is present?

Given that the results presented here and found in previous work (Knobe, 2003a, 2006, 2010) suggest that there is an influence of normative considerations on social-cognitive judgments are we required to drastically revise of our understanding of social cognition to accommodate an evaluative moral function as some have suggested (Knobe, 2010), or can we account for this normative influence while largely preserving our models of social cognition? The findings of the set of studies presented in chapter two strongly suggest a potential bi-directional relationship, based on the inferences that norm-violation allows for. This relationship allows for a moral influence but would not require a drastic revision of social cognition, as the basic process and function remain the same. Results from this first set of studies presented show that the side-effect effect remains when there is only a difference in norm-violation, but not in moral status. The side-effect effect also remained when the norms that were used were conventional, rather than moral. Finally, the results show that in addition to judgments of intention, predictions of future action are impacted. These results suggest that the side-effect effect, an example of moral understanding impacting social-cognitive judgments, is driven by differences in norms and is consistent with the main function of explaining and predicting behavior. Norms influence mental state ascriptions and cause a side-effect effect because norm-violating behavior supports the ascription of counternormative mental states, which in turn influence ascriptions of intentional action, predictions of future behavior, and other judgments relevant to Theory of Mind. This interpretation maintains the explaining, predicting and controlling function of social cognition and incorporates the influence of moral understanding without viewing social-cognitive capacities as biased, broken, or primarily evaluative.

The results of the studies concerning explanation discussed in chapter three and four also support the argument that a drastic revision is not necessary. In study two we saw participants providing Alternative Causal History explanations which suggests that people still view norms as having a causal influence on mental states, but that there does not need to be an explicit belief present. Instead they see other paths for the influence of norms on behavior. That norms seem to influence independent of belief does not indicate

norms function outside of typical Theory of Mind, but instead suggests explicit belief is not the only acceptable causally explanatory pathway. In the third set of studies the use of norms in ethical explanations for socio-historical change is related to people's meta-ethical beliefs. This relationship suggests that, at least in this instance, what makes norms explanatory is having particular views about the nature of the moral world that may give moral norms qualities, perhaps an objective causal directed force, that allow norms to function in a similar way to other candidate explanations. These two sets of studies provide evidence that norms are used as explanations because they capture something about how the world is, not just how it ought to be. They suggest that norms are viewed as an important source of information concerning how, what, and why events might happen.

Now that I have argued that the influence of moral understanding is the result of the normative qualities rather than only the moral aspects of moral understanding, the next logical question to be asked is, are there are any differences in the influence of norms depending on norm types? While the evidence presented in the preceding chapters suggests that all sorts of norms can serve as a source of information, these studies also present evidence to suggest that there may be some differences between the different types of norms. In chapter three the results specifically indicate differences between how different norms are used as explanations as well as how they impact other types of explanations and relate to other sources of information. In particular, we see evidence that moral norms may be different in their relationship to belief. When norms were moral, participants were less sensitive to belief and action matching when producing norm explanations and rating belief explanations, and more sensitive when both producing and rating preference explanations. Additionally for moral norms, whether or not the agent's action matched the norm of the culture did not impact the production of Alternative Causal History explanations. All this perhaps indicates that moral norm explanations are accepted and provided in the absence of explicitly worked out mental states mediating the relationship between norms and actions.

The set of studies in chapter four show that differences in how people conceptualize the moral world lead to differences in the explanations people endorse for social-historical change. These results suggest that the way people think about the nature of a norm impacts what information a norm can provide, and thus what role a norm can play in social-cognitive judgments. These results from chapter four, along with the results of chapter three, give good reason to think that if different types norms are grounded in different ways, for example moral norms being universal concerns of harm versus conventional norms appealing to rules or authority (Turiel, 1983), than different norms will interact with social-cognitive judgments in different ways. The information a norm can provide depends on what people believe about the norm, and this difference in information impacts the possible inferences or explanations, as the norms can only impact these judgments when there are plausible mediating paths.

Though the set of studies in chapter two was designed to demonstrate that the side-effect effect was about norm-violation and not moral status, there is some reason to see the results as indicating the extent to which a norm is seen as shared will affect its impact on judgments of intention. In experiment 1 conventional norms and moral norms aligned with our own (i.e. do not harm) were created for the purposes of the study and simply presented to participants. These norms produced the side-effect effect similar to

established norms. However, in experiment 2 we attempted to pull apart norm status and moral status by establishing a norm for a scenario that ran counter to the moral norm held by participants (i.e. do not harm). Establishing this counter norm for a sub-group required a much more elaborate presentation. Even with this presentation the side-effect effect was eliminated, but not completely reversed. That it was very easy to establish conventional and consistent moral norms, but difficult to establish a moral norm counter to our own for a sub-group, supports the well established idea that there may be differences between how shared people believe norms are (Turiel, 1983.)

While we have demonstrated that other types of norms do impact social-cognitive judgments, there is some evidence to suggest that moral norms are unique. This may be because all norms are a source of information, but they might not all provide the same sort of information.

What then is the view of the relationship between social cognition and moral understanding I wish to defend here? I wish to argue that social cognition and moral understanding have a bidirectional relationship. In addition to the accepted view of the relationship where our social cognitive capacities produce information that can be used for our moral understanding (Cushman and Young, 2011; Kohlberg, 1969; Malle and Nelson, 2003), our moral understanding, specifically the normative reasoning capacities, can provide information that is utilized in social cognitive judgments. Our normative reasoning capacities, typically employed in making evaluative judgments about what should, ought to or typically does happen, can because norms provide a "desire independent reason for action" (Searle, 2001) be used as a source of information that is then helpful in the social-cognitive process predicting or explaining behavior.

However, it is important to note that the view presented here holds that the bi-directional relationship between moral understanding and social cognition exists as a result of these capacities performing their main functions of evaluating, or explaining and predicting. The fact that there is a normative standard, which is necessary for evaluation, along with the understanding or belief that others are made aware of or are otherwise exposed to this standard, creates an expectation that can be used to explain and predict as well as a pivot point from which to infer mental states that can again be used to explain and predict. This information is used in social-cognitive judgments in much the same way as other information is used in these judgments, with the main difference being the source. In this way the relationship between social-cognition and moral understanding is not bi-directional for any specific moral reason, but because of the normative quality of morality. This view separates itself from a few others in that it maintains the traditional role of social cognition (Knobe, 2010) and holds neither capacity as subordinate (Wellman and Miller, 2008.) The view presented in this dissertation suggests that in accomplishing separate functions these two capacities each provide for the other a source of information that the other capacity can use to better accomplish its function.

While I have presented evidence in support of this view there are still many open questions concerning the relationship between social cognition and moral understanding. Though all three studies in some way address the issue there are still many open questions about what the differences between norm types mean for the relationship with social cognition. I have shown that it appears there are some differences in the relationship when the type of norm varies, specifically suggesting that moral norms seem to be different in some ways. I have suggested that if beliefs concerning the norm differ,

what can be inferred from these norms may differ. Research that explicitly examines more of the possible differences between norms and different attributions may be necessary to make strong claims about how different norm types interact with social cognitive judgments.

Open questions also remain concerning how exactly people represent the path of norm influence when it is not mediated by beliefs. We have presented some evidence for alternative causal history explanations, where the influence of norms is mediated but not by beliefs. Questions remain concerning exactly what representations are supporting these explanations, the extent to which these representations are explicitly held, and the extent to which these representations play a role in other social-cognitive judgments.

These open questions and other issues present some logical future directions. In particular I would like to expand on the work concerning norms as explanations to address questions concerning possible differences in how the actions of individuals, groups, or aggregate groups are interpreted and understood. The two studies presented here look at explaining the actions of an individual, and explaining why something involving an aggregate group occurred, but given the work involving differences between explanations for these different group types (O'Laughlin and Malle, 2002) examining these different groups in a more systematic way promises to be interesting.

Another direction for further research is work that examines the relationship between the type of explanation used to explain the behavior of agents or groups and whether or not people judge the behavior to be best described by prescriptive or descriptive norms. Past work on explanation has made distinctions between different types of explanations (Keil, 2006; Lombrozo, 2006, 2010.) For example, one could give an explanation at a mechanistic level (explain a pen by the manufacturing process that brought it about), or a functional level (explain a pen by its function of leaving marks on paper), or an intentional level (explain a pen by the goal of writing). Perhaps there is a relationship between the level of explanation used and the norms people may infer. If this is true in borderline cases might introducing a type of behavior or action by using teleological explanations rather than intentional or mechanistic lead people to view the behavior as conventional rather than moral or descriptive?

There is also much ground to be covered in investigating exactly what the alternative paths from norms to behavior are. While chapter three argued for non-belief mediated paths, the exact mechanics of the alternatives remain open. A study that more systematically examines the potential alternatives, including those found in the second experiment of chapter three, should provide further insight into the relationship between norms and explanation.

In conclusion I have argued that indeed the best way to describe the relationship between social cognition and moral understanding is as bi-directional. However, this change in the description of this relationship does not change how we should understand the primary roles of either capacity. Additionally, I have argued that the reason for the bi-directionality is due to the normative aspects of our moral understanding, but that there are differences between norms that may make moral norms unique in many respects of the relationship.

Chapter 6: References

Adams, F. and Steadman, A. (2004a). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, 64, 173-181.

Ahn, W. & Kalish, C. (2000). The role of mechanism in causal reasoning. In R. Wilson, & F. Keil (Eds.) *Cognition and explanation*, Cambridge, MA: MIT Press.

Apperly, I.A., Riggs, K.J., Simpson, A., Chiavarino, C. and Samson, D. (2006). Is belief reasoning automatic? *Psychological Science,* 17, 841-844.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.

Blassingame, J. W. (Ed.) (1977). *Slave testimony: Two centuries of letters, speeches, interviews, and autobiographies.* Baton Rouge, LA: Louisiana State University Press.

Brink, David (1989). *Moral Realism and the Foundation of Ethics*. Cambridge: Cambridge Unviersity Press.

Cal. Penal Code § 187

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: M. I. T. Press.

Cohen, J. (1997). The arc of the moral universe. *Philosophy and Public Affairs,* 91-134.

Craver, C. (2007). *Explaining the brain: What a science of the mind-brain could be*. New York: Oxford University Press.

Cusman, F. & Young, L. (2011) Patterns of moral judgment derive from non-moral psychological representations. *Cognitive Science,* 35, 1052-1075.

Dalbert, C., Montada, L., & Schmitt, M. (1987). Glaube an eine ge- rechte welt als motiv: Vali-dierungskorrelate zweier Skalen [Belief in a just world as motive: Validity correlates of two scales]. *Psychologische Beitrage*, *29*, 596–615.

Darden, L. (2006). *Reasoning in biological discoveries: Essays on mechanisms, interfiled relations, and anomaly resolution*. Cambridge: Cambridge University Press.

Darley, J. M., & Shultz, T. R. (1990). Moral Rules - Their Content And Acquisition. Annual Review of Psychology, 41, 525- 556.

Fincham, F. D., & Jaspers, J. (1979). Attribution Of Responsibility To The Self And Other In Children And Adults. Journal of Personality and Social Psychology, 37(9), 1589-1602.

Fiske, S.T. (1989). Examining the role of intent. In J.S. Uleman & J.A Bargh (Eds). *Unintended Thought* (p 253-283) New York, NY: The Guilford Press.

Fiske, S.T., & Taylor, S.E. (2007). *Social Cognition: From brains to Culture.* New York, McGraw-Hill Higher Education.

Fodor, J.A. (1992). A theory of the child's theory of mind. *Cognition, 44,* 283-296.

Furnham, A. (2003). Belief in a just world: research progress over the past decade. *Personality and Individual Differences, 34*, 795-817.

Gopnik, A. (1999). Theory of mind. In R. Wilson & F. Keil (Eds*). The MIT Encyclopedia of the Cognitive Sciences* (p. 838) Cambridge, MA: M. I. T. Press.

Goodwin, G. P. & Darley, J. M. (2008). The psychology of meta-ethics: Exploring objectivism. *Cognition, 106,* 1339-1366.

Guglielmo, S. and Malle, B. F. (2008) The Knobe artifact? Lessons in the subtelties of language. Pre-conference workshop on Experimental Philosophy, Society of Philosophy and Psychology 34th Annual Meeting, Philadelphia, PA.
Guglielmo, S. and Malle, B. F. (2009a) Can unintended side--effects be intentional? Solving a puzzle in people's judgments of morality and intentionality. Submitted for publication.
Haidt, J. & Joseph, C. (2004) Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. Daedalus, 4, 55-66.
Harman, G. (1975). Moral relativism defended. *Philosophical Review, 84,* 3–22.
Harman, G. (1977). *The nature of morality*. Oxford: Oxford University Press.
Heider, F. (1958). The psychology of interpersonal relations. New York,: Wiley.
Hitchcock, C. & Knobe, J. (2009). Cause and Norm. *Journal of Philosophy, 106,* 587-612.
Holton, R. (2010). Norms and the Knobe effect. Forthcoming in *Analysis.*
Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219-266). New York: Academic Press.
James, W. (1981) *The principles of psychology.* (2 vols.) Cambridge, MA: Harvard University Press. (Original work published 1890)
King, M. L., Jr., (1986). *A Testament of hope: The Essential writings of Martin Luther King, Jr.* In J. Washington (Ed.), San Francisco: Harper and Row.
Kalish, C.W. (1998). Reasons and causes: Children's understanding of conformity to social rules and physical laws. *Child Development*, 69, 706-720.
Kalish, C.W. (2006). Integrating normative and psychological knowledge: What should we be thinking about? *Journal of Cognition and Culture*, 6, 191-208.
Kalish, C. W., & Cornelius, R. (2007). What is to be done? Children's ascriptions of conventional obligations. *Child Development*, 78, 859-878.
Kalish, C.W. & Shiverick, S.M. (2004). Children's reasoning about norms and traits as motives for behavior. *Cognitive Development*, 19, 401-416.
Keil, F.C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Science, 7,* 368-373.
Keil, F.C. (2006). Explanation and Understanding. Annual Review of Psychology. 57, 227-254.
Kelley, H.H. (1967) Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15 pp. 192-240). Lincoln: University of Nebraska Press.
Kelly, D., Stich, S., Haley, K.J., Eng, S.J., & Fessler, D. M. (2007). Harm, affect, and the conventional/moral distinction. *Mind and Language, 22*(2), 117-131.
Knobe, J. (2003a). Intentional action and side effects in ordinary language. Analysis, 63, 190–193.
Knobe, J. (2003b). Intentional action in folk psychology: An Experimental investigation. *Philosophical Psychology*, *16*, 309-324.
Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, *64*, 181-187.
Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, 130, 203-231.

Knobe, J. (2007). Reason explanation in folk psychology. *Midwest Studies in Philosophy*. 31, 90-107.
Knobe J. (2010). Person as scientist, person as moralist. Behavioral and Brain Sciences, 33, 315-365.
Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H., & Sommers, T. (2012). Experimental philosophy. *Annual Review of Psychology, 63,* 81-99.
Knobe, J. & Burra, A. (2006). Intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition 6*, 113-132.
Knobe, J. & Mendlow, G. (2004). The good, the bad, and the blameworthy: Understanding the role of evaluative considerations in folk psychology. *Journal of Theoretical and Philosophical Psychology 24*, 252-258.
Kohlberg, L. (1981/1969). Essays on Moral Development, Volume1: The Philosophy of Moral Development. New York: Harper Row.
Kuhn, D. (2001). How do people know? *Psychological Science, 12,* 1-8.
Leiter, B. (2001). Moral facts and best explanations. *Social Philosophy and Policy, 18*, 79-101.
Lerner, M. (1980). *The belief in a just world: a fundamental delusion*. New York: Plenum.
Leslie, A., Knobe, J. & Cohen, A. (2006). Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment. *Psychological Science, 17*, 421-427.
Lin, S., Keysar, B., & Epley, N. (2010) Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Cognition,* 46, 551-556.
Lombrozo, T. (2006). The structure and function of explanations. Trends in Cognitive Sciences, 10, 464-470.
Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science,* 33, 273-286.
Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. Cognitive Psychology, 61, 303-332.
Lombrozo, T. (in press). Explanation and abductive inference. K.J. Holyoak and R.G. Morrison (Eds.), *Oxford Handbook of Thinking and Reasoning*.
Lombrozo, T. & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99*, 167-204.
Lucas, C., Griffiths, T. L., Xu, F., & Fawcett, C. (in press). A rational model of preference learning and choice prediction by children. *Advances in Neural Information Processing Systems*, 21.
Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language,* 23, 165-189.
Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3,* 23-48.
Malle, B. F. (2004). How the mind explains behavior: Folk explanations, meaning, and social interaction. Cambridge, MA: MIT Press.
Malle, B. F. & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, 33, 101-121.
Malle, B. F., & Nelson, S. E. (2003). Judging Mens Rea: The Tension Between Folk Concepts and Legal Concepts of Intentionality. *Behavioral Sciences and the Law, 21*, 563-580.

Mallon, R. (2008). Knobe vs. Machery: Testing the trade-off hypothesis. *Mind & Language*, 23, 247-255.
Mele, A. (2001). Acting intentionally: Probing folk notions. In B. F., Malle, L. J. Moses, & D. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: M. I. T. Press.
Mele, A.R. & Cushman, F. (2007). "Intentional action, folk judgments, and stories: Sorting things out." *Midwest Studies in Philosophy*, 31, 184-201.
Miller, G.A., Galanter, E., and Pribram, K.H. (1960) *Plans and the structure of behavior.* New York: Holt, Rinehart and Winston.
Nadelhoffer, T. (2004b). Praise, side effects, and intentional action. *The Journal of Theoretical and Philosophical Psychology*, *24*, 196-213.
Nichols, S. (2004). After objectivity: An empirical study of moral judgment. *Philosophical Psychology, 17,* 5-28.
O'Laughlin, M. J. and Malle, B. F. (2002). How people explain actions performed by groups and individuals. *Journal of Personality and Social Psychology,* 82, 22-48.
Pettit, D. and Knobe, J. (2009) The pervasive impact of moral judgment. *Mind and Language*, 24, 586-604.
Phelan, M. and Sarkissian, H. (2006). The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, 138, 291-298.
Piaget, J. (1965/1932). The Moral Judgment of the Child. New York: Free Press.
Rozenblit, L.R. & Keil, F.C. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science, 26,* 521-562.
Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
Sarkissian, H., Park, J., Tien, D., Wright, J.C., & Knobe, J. (2011). Folk moral relativism. *Mind & Language, 26,* 482-505.
Searle, J.R. (2001). Rationality in action. Cambridge, MA: M. I. T. Press.
Sheehy, E., Conrad, S.L., Brigham, L.E., Luskin, R., Weber, P., Eakin, M. et al. (2003) Estimating the number of potential organ donors in the united states. *New England Journal of Medicine*, 349, 667-674.
Shtulman, A. (2010). Theories of God: Explanatory coherence in a non-scientific domain. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (1295-1300). Austin, TX: Cognitive Science Society.
Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of Moral Responsibility and Punishment. Child Development, 57(1), 177-184.
Sinnott-Armstrong, Walter, "Moral skepticism", *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2009/entries/skepticism-moral/>.
Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development, 52,* 1333 – 1336.
Skowronski, J. J., and Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. Psychological Bulletin, 105, 131-142.

Sloman, S.A., Fernbach, P.M., and Ewing, S. (2010). A Causal Model of Intentionality Judgment. Submitted for publication.
Sripada, C. S. (2009). The Deep Self Model and Asymmetries in Folk Judgments About Intentional Action. Philosophical Studies. In Press.
Sripada, C.S. and Konrath, S. (2010). Telling more than we can know about intentional action. Submitted for publication.
Thagard, P. (1989). Explanatory Coherence. *Behavioral and Brain Sciences, 12,* 435-502.
Tolman, E.C. (1925). Purpose and Cognition: The determiners of animal learning. *Psychological Review*, 32, 285-297.
Turiel, E. (1983). The development of social knowledge: Morality and convention. New York, NY; Cambridge University Press.
Uttich, K. & Lombrozo, T. (2010) Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116,* 87-100.
Uttich, K., Tsai, G. & Lombrozo, T. (In preparation) Ethical explanations, moral objectivity, and moral progress. Manuscript in preparation.
Wainryb, C., Shaw, L., Langley, M., Cottam, K. and Lewis, R. (2004). Children's thinking about diversity of belief in the early school years: judgments of relativism, tolerance and disagreeing persons. *Child Development*, *75*, 687–703.
Weiner, B. (1995). Judgments of responsibility : A foundation for a theory of social conduct. New York: Guilford Press.
Wellman, H.M. (1992). The Child's Theory of Mind. Cambridge, MA: M. I. T. Press.
Wellman, H.M. and Miller, J.G. (2006) Developing conceptions of responsive intentional agents. *Journal of Cognition and Culture*, 6, 27-55.
Wellman, H.M. & Miller, J.G (2008). Including deontic reasoning as fundamental to theory of mind. Human Development, 51, 105-135.
Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13,* 103-128.
Williams, B. 1985: *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.
Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
Woodward, J. (2003). Counterfactuals and causal explanation. International Studies in the Philosophy of Science, 18, 41-72.
Woodward, James, "Scientific Explanation", The Stanford Encyclopedia of Philosophy (Spring 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/scientific-explanation/>.
Wright, J.C. & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind & Language, 24*, 24-50.
Wright, L. (1976). *Teleological explanations*. Berkeley, CA: University of California Press.