Cognition 116 (2010) 87–100

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Norms inform mental state ascriptions: A rational explanation for the side-effect effect

Kevin Uttich *, Tania Lombrozo

University of California, Berkeley, United States

ARTICLE INFO

Article history:
Received 11 April 2009
Revised 6 April 2010
Accepted 6 April 2010

Keywords:
Social cognition
Side-effect effect
Knobe effect
Morality
Moral Psychology
Theory of mind
Intentional action
Norms
Intentionality

ABSTRACT

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.

1. 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 need not 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 base rates. These observations suggest that norms *should* inform mental state ascriptions if reasoners are to be effective "intuitive scien-

* Corresponding author.
E-mail address: uttich@berkeley.edu (K. Uttich).

0010-0277/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2010.04.003

tists" (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 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 & 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 & Mendlow, 2004; Machery, 2008), across cultures (Knobe & Burra, 2006), and with preschool children (Leslie, Knobe, & Cohen, 2006). While a variety of explanations for the effect have been offered (Adams & Steadman, 2004; Knobe, 2006; Machery, 2008; Nadelhoffer, 2004), no single proposal successfully accounts for all the data collected to date (Pettit & 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 & Steadman, 2004), the desire to blame an agent for a negative outcome (Malle & Nelson, 2003; Mele, 2001), or an emotional reaction (Nadelhoffer, 2004) 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 counternormative 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 Malle and Knobe (1997) model of intentional action (for related arguments about differences in desire across conditions see Guglielmo & Malle, submitted for publication; see also Sripada, in press, 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 & Davis, 1965), the cue-diagnosticity approach to trait attribution (Skowronski & Carlston, 1987), and the Covariation ANOVA Model (Kelley, 1967), many of which emphasize the importance of atypical (and hence counter-normative) behavior in guiding judgment (see also Malle & Guglielmo, 2008; Holton, 2010; Sripada, in press; Sripada & Konrath, submitted for publication).

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 norms.

2. 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*).[1] While norm status varies across 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.

2.1. 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.

2.2. 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

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

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 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 1–7 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.

2.3. Results and discussion

Participants' ratings of whether it is appropriate to say that the agent brought about the side effect "intentionally" (see Fig. 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 aver-

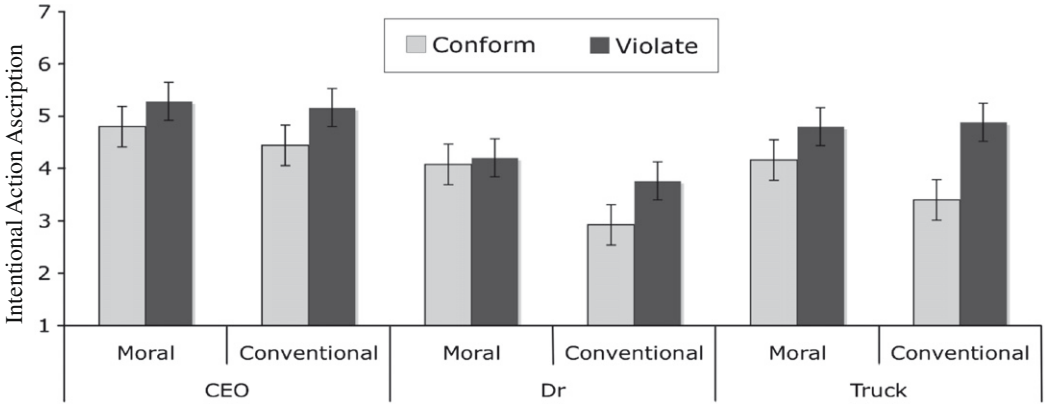

Fig. 1. Experiment 1 ratings of how appropriate it is to describe an action as having been performed intentionally as a function of norm status and norm type. Ratings were made on a scale from 1 (neither appropriate nor inappropriate to say outcome brought about intentionally) to 7 (appropriate to say outcome brought about intentionally) with 4 (neither appropriate nor inappropriate) as a midpoint.

age ratings for the norm-violating side effects were numerically higher than those for the norm-conforming side effect.[2]

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

3. 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.

3.1. 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 eight participants in each of 12 conditions.

3.2. 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,

[2] 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; $t(429) = -3.28$, $p < .01$).

Table 1
Judgments from Experiment 2 as a function of context and side effect valence. Means are followed in parentheses by standard deviations. The patterns of shading highlight significant differences across conditions. Main effects and interactions are also indicated in the right-hand portion of the table, with a single asterisk (*) indicating a significant effect at the $p < .05$ level, and a double asterisk (**) indicating a significant effect at the $p < .01$ level.

Question	Superhero		Supervillain		Main effects		Interactions		
	Good SE	Bad SE	Good SE	Bad SE	SE valence	Context	SE valence × Context	Context × 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)	**	**			*

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 not to 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 1–7 scale as in Experiment 1. Participants made additional judgments (see Table 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).

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).

3.3. 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. 2). There were no other significant effects.

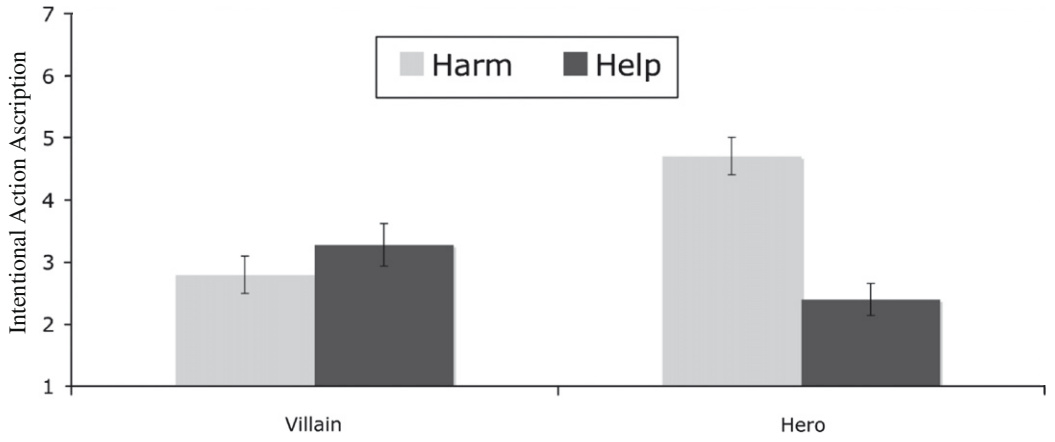

Fig. 2. Experiment 2 ratings of how appropriate it is to describe an action as having been performed intentionally as a function of norm status and context. Presented on a scale from 1 (not appropriate to say outcome brought about intentionally) to 7 (appropriate to say outcome brought about intentionally),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

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, 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 and 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, and 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 and 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 of 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 for 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 from 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 agents 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 participant's 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 moral 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 out-

comes 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.

4. 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.

4.1. Participants

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

4.2. 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]

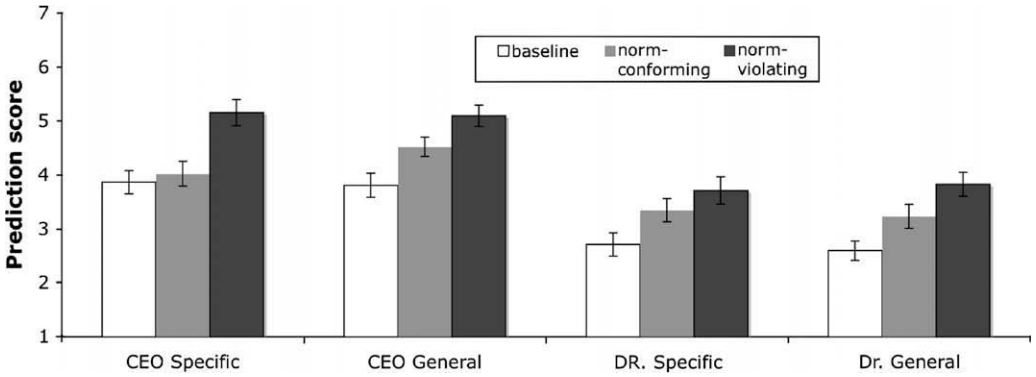

Fig. 3. Prediction scores from Experiment 3 on a scale from 1 (good side effect likely in future) to 7 (bad side effect likely in future).

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.

4.3. 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. 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 three-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: $t(102) = 1.91$, $p = .059$).[3] 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 partici-

[3] 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 ($t(34) = 5.13$, $p < .001$).

pants 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 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.

5. 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 (Experiments 1 and 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, 1890/1981; Miller, Galanter, & Pribram, 1960; Tolman, 1925). 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.

5.1. Relationship to previous accounts

While other accounts of the side-effect effect can be modified to accommodate our findings, the Rational Scien-

tist 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, 2003a, 2003b, 2006; Nadelhoffer, 2004; Wright & Bengson, 2009). However, subsequent studies using similar vignettes have produced examples that counter these accounts. For example, Phelan and Sarkissian (2006) 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 & 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 & Malle, submitted for publication.)

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 does not 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 2×2 design used in our Experiment 2 (superhero or supervillain context $\times$ 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, & Ewing, submitted for publication; see also Guglielmo & Malle, submitted for publication, and Sripada, in press, 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.

5.2. 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 Kalish, 2006; Wellman & Miller, 2006, 2008). 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, Riggs, Simpson, Chiavarino, & Samson, 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, Griffiths, Xu, & Fawcett, 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.

6. Conclusion

While we have contested Knobe's (2003a, 2003b, 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.

Acknowledgments

We thank Joshua Knobe and Edouard Machery for comments on an earlier draft, Lori Markson, Jennifer Cole Wright, Bertram Malle, and Steven Sloman for relevant conversations, and the Berkeley Moral Psychology group and Child Cognition Lab for helpful feedback. Finally, we'd like to thank Jesse van Fleet and the other members of the Concepts and Cognition lab for feedback and help with data collection.

References

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

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

Cal. Penal Code, Section 187.

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

Gopnik, A. (1999). Theory of mind. In R. Wilson & F. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 838). Cambridge, MA: M.I.T. Press.

Guglielmo, S., & Malle, B. F. (submitted for publication). Can unintended side-effects be intentional? Solving a puzzle in people's judgments of morality and intentionality.

Holton, R. (2010). Norms and the Knobe effect. Forthcoming in *Analysis*.

James, W. (1890/1981). *The principles of psychology (2 vols.)*. Cambridge, MA: Harvard University Press [Original work published, 1890].

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.

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.

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.

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., & 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.

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.

Lucas, C., Griffiths, T. L., Xu, F., & Fawcett, C. (2009). 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., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101–121.

Malle, B. F., & Guglielmo, S. (2008). The Knobe artifact? Lessons in the subtleties of language. In *Pre-conference workshop on experimental philosophy, society of philosophy and psychology 34th annual meeting*, Philadelphia, PA.

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., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart and Winston.

Nadelhoffer, T. (2004). Praise, side effects, and intentional action. *The Journal of Theoretical and Philosophical Psychology, 24*, 196–213.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind and Language, 24*, 586–604.

Phelan, M., & 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.

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.

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52*, 689–699.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (submitted for publication). A causal model of intentionality judgment.

Sripada, C. S. (in press). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies*.

Sripada, C. S., & Konrath, S. (submitted for publication). Telling more than we can know about intentional action.

Tolman, E. C. (1925). Purpose and cognition: The determiners of animal learning. *Psychological Review, 32*, 285–297.

Wellman, H. M. (1992). *The child's theory of mind*. Cambridge, MA: M.I.T. Press.

Wellman, H. M., & 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.

Wright, J. C., & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind & Language, 24*, 24–50.