The Centrality of Belief and Reflection in Knobe Effect Cases: A Unified Account of the Data

Mark Alfano (University of Oregon), James R. Beebe (University at Buffalo)
and Brian Robinson (Grand Valley State University)

Forthcoming in *The Monist* 95 (April 2012), special issue on experimental philosophy

Recent work in experimental philosophy has shown that people are more likely to attribute intentionality, knowledge, and other psychological properties to someone who causes a bad side-effect than to someone who causes a good one. We argue that all of these asymmetries can be explained in terms of a single underlying asymmetry involving belief attribution because the belief that one's action would result in a certain side-effect is a necessary component of each of the psychological attitudes in question. We argue further that this belief-attribution asymmetry is rational because it mirrors a belief-formation asymmetry and that the belief-formation asymmetry is also rational because it is more useful to form some beliefs than others.

Keywords: Knobe effect, experimental philosophy, belief, heuristics

1. Introduction

In his seminal paper on the side-effect effect, Joshua Knobe (2003a) identified an asymmetry in the experimentally measured attribution of intentionality to the side-effects of actions. In the experiment's HARM condition, participants read a vignette in which a corporate chairman was told that by instituting a new policy he would raise profits and (as a side-effect) harm the environment. In the HELP condition, participants read an otherwise identical vignette in which the new policy would help the environment. In both vignettes, the chairman responded that he cared only about profit, not the environment. The new policy was instituted; profits increased; the environment was affected as predicted. Participants were then asked to say whether the chairman *intentionally* brought about this side-effect. The asymmetry showed up in their responses: more people said that he intentionally harmed than said he intentionally helped.

Since 2003, an entire literature has sprung up around this phenomenon. Experiments conducted for follow-up articles, responses, responses to follow-ups, and responses to responses have generated a wealth of further data. These experiments have greatly expanded the scope of the cases in which the asymmetry has been detected. Knobe's original study was meant to connect *morally bad* outcomes with attributions of *intentionality*. Later experiments have varied both the type of outcome and the property attributed. For instance, Thomas Nadelhoffer (2006) and Knobe and Mendlow (2004) found the same asymmetry when the outcome was not morally bad/good but *prudentially bad/good* (viz. by helping another person, the protagonist of the vignette decreased his own well-being). Knobe (2004a) discovered the same asymmetry when the outcome was *aesthetically bad/good*, and again (2007) when it was *in violation of/conformity*

to a bad law (in a *Schindler's List*-style case, where a greedy factory owner violates/conforms to a racial identification law akin to the one used in Nazi Germany).

Other studies have found the same asymmetry when the attribution had to do not with *intentionally* bringing about a side-effect but with other psychological properties. For instance, Knobe (2004b, 2006) showed that people exhibit the same attribution asymmetry when asked whether the protagonist brought about a bad/good side-effect *in order to* achieve his main goal. Pettit and Knobe (2009) noted the same effect when people were asked to say whether someone *decided* to bring about the side-effect, was *in favor of* bringing about the side-effect, or *advocated* bringing about the side-effect. Tannenbaum, Ditto and Pizarro (2007) report the same asymmetry in judgments about whether the protagonist *desired* to bring about the side-effect. Perhaps the most surprising results, however, are due to Beebe and Buckwalter (2010), Beebe and Jensen (forthcoming), and Beebe (forthcoming a) who showed that the asymmetry crops up when participants are asked whether the protagonist *knew*, *believed,* or *should have believed* that his action would bring about the side-effect. Almost every existing explanation of the general class of side-effect effects assumes that participants interpret the chairman as both believing with equal confidence and knowing that the side-effect will occur in the help and the harm conditions. The recent studies of Beebe and his collaborators, however, invalidate many of these explanations. These results also form the basis for the explanation that we offer, which is that respondents are responding appropriately to the vignettes because they rationally judge protagonists to have formed stronger beliefs about the side-effects in harmful or norm-violating cases than in helpful or norm-conforming cases.

2. Doxastic Heuristics

We propose that all of the existing data on the side-effect effects are best explained in terms of a belief-formation heuristic. Consider the following set of claims:

(Know $\rightarrow$ Believe) Agent *a* knows that φ'ing would make it the case that *p* only if *a* believes that φ'ing would help to make it the case that *p*.

(Intentionally $\rightarrow$ Believe) Agent *a* intentionally makes it the case that *p* by φ'ing only if *a* believes that φ'ing would help to make it the case that *p*.

(In order to $\rightarrow$ Believe) Agent *a* φ's in order to make it the case that *p* only if *a* believes that φ'ing would help to make it the case that *p*.

(Desire $\rightarrow$ Believe) Agent *a* desires to make it the case that *p* by φ'ing only if *a* believes that φ'ing would help to make it the case that *p*.

(Decide $\rightarrow$ Believe) Agent *a* decides to make it the case that *p* by φ'ing only if *a* believes that φ'ing would help to make it the case that *p*.

(In favor of $\rightarrow$ Believe) Agent *a* is in favor of making it the case that *p* by φ'ing only if *a* believes that φ'ing would help to make it the case that *p*.

(Advocate $\rightarrow$ Believe) Agent *a* advocates making it the case that *p* by φing only if *a* believes that φ'ing would help to make it the case that *p*.

Each of these claims seems intuitively plausible. (Know $\rightarrow$ Believe), for instance, is almost universally accepted in epistemology. To refute (Intentionally $\rightarrow$ Believe), one would have to find a case where someone intentionally brings about an effect despite his not believing that his actions will produce the effect. Suppose, for instance, that Jacob rolls a fair die, believing that it

will not come up 6 because there is only a 17% chance of rolling a 6 with a fair die. Despite the low probability, he does indeed roll a 6. Does John intentionally roll a 6 without believing he will? It seems clear that the answer is negative. Just as one cannot intentionally win a fair lottery, though one can intentionally put oneself in a position where it is possible to win a fair lottery by buying a ticket, so one cannot intentionally roll a 6 on a fair die. At the very least, it seems conceptually impossible to intentionally bring about an effect by performing an action without dispositionally believing that the action raises the probability of the effect.

To refute (In order to $\rightarrow$ Believe), one would have to find an analogous case where someone does not believe his action will produce an effect, and yet he manages to perform the action in order to bring about the effect. Again, we see very little prospect of finding such a case.

To refute (Desire $\rightarrow$ Believe), one would have to find a case where someone desires to bring about an effect by means of her action yet does not believe that her action will bring about the effect. Imagine Jane, who has locked herself out of her apartment, landing blow after blow upon the front door in an attempt to break it down. When her neighbor asks what she is doing, she responds, 'I want to knock down this door, so I'm kicking it and hitting it.' Her neighbor looks at her quizzically and reminds her that the lock and hinges are made of stainless steel. Jane says, 'Yes, I understand. I won't knock down the door by doing this, but I still want to knock down the door this way.' Perhaps such a case is possible, but only a patently irrational agent would have such a combination of beliefs and desires. The irrationality argument may also apply to (Decide $\rightarrow$ Believe), (In favor of $\rightarrow$ Believe), and (Advocate $\rightarrow$ Believe). Jane has decided to knock down her sturdy door by kicking it (is in favor of knocking down the door by kicking it / advocates knocking down the door by kicking it), despite the fact that she has no confidence that

she will succeed. Again, though, only a thoroughly irrational agent could make such a decision, have such a pro-attitude, or advocate such a project.

Even if the conditionals above are not logically true or exceptionless, they surely hold in a great majority of actual cases. That alone should be sufficient to ground our arguments in this paper.

If the foregoing reflections are correct, belief might be the unifying factor in all of the side-effect studies conducted to date. We hypothesize that all of the protagonists that appear in the various vignettes used in the literature have practical reasons for giving greater attention to and engaging in deeper reflection about the potential side-effects of their actions in one condition than another. This greater attention and reflection makes the agents more likely to form beliefs about the side-effects in question and also makes attributors more likely to interpret them as forming such beliefs. A common instance of this type of situation occurs when agents are thinking about violating a salient norm. If people attribute belief when a side-effect violates a norm, but not when it conforms to a norm, they can consistently affirm or deny knowledge, intentionality, acting in order to, desire, decision, being in favor of, and advocating of a bad/good side-effect simply by applying *modus tollens*. No belief? Then no knowledge, no intention, no acting in order to, no desire, no decision, no being in favor of, and no advocating. For any relevant psychological property *X*, people tend to attribute *X* in norm-violation conditions but tend not to attribute it in norm-conformity conditions because they attribute belief in the norm-violation conditions but not in the norm-conformity conditions. Let us call this the *Norm-violation / Belief-attribution heuristic*:

(NBA) If another agent *a*'s φ'ing would make it the case that *p* and *p* violates a norm salient to *a*, then attribute to *a* the belief that φ'ing would make it the case that *p*.[1]

(NBA)'s widespread acceptance would unify the asymmetrical attributions but leave them mysterious. Even if it is granted that (NBA) explains the asymmetry of people's attribution of psychological states, it remains unclear why belief attributions are subject to this asymmetry.

Fortunately, belief attributions not only unify the phenomena but also explain them. We need not regard those who use (NBA) as irrational belief-attributers. We contend that (NBA) is rational because people also employ the following *Norm-violation / Belief-formation heuristic*:

(NBF) If my own φ'ing would make it the case that *p* and *p* violates a norm salient to me, believe that φ'ing would make it the case that *p*.[2]

If (NBF) is correct, then the attribution asymmetry generated by (NBA) tracks a real belief-formation asymmetry. This takes the pressure off (NBA). After all, it is rational to attribute beliefs to people who violate norms (and not to attribute beliefs to people who conform to them) if those very people form beliefs when they violate norms but not when they conform. However, (NBF) shifts the pressure onto itself: why would people form beliefs about the effects of their own actions in this way? Any theory that entails widespread irrationality is *prima facie* implausible, so we need to argue that employing (NBF) is not irrational.

To see why (NBA) and (NBF) are both rational heuristics, consider the fact that, though true beliefs are typically worth having, some true beliefs are more worth having than others. For instance, the true belief that *n* is the winning number in a lottery is more valuable than the true belief that *m* is not the winning number. What makes (NBF) plausible is the fact that true beliefs

[1] Note that (NBA) is not formulated as a biconditional. We do not think its converse is true. However, we do think that people fail to form beliefs about norm conformity for which they have evidence at a much higher rate than they fail to form beliefs about norm violation for which they have evidence.

[2] We are not claiming that people consciously think in this way. Rather that they form their own beliefs according to this non-conscious heuristic.

to the effect that one is violating a norm are typically more valuable than true beliefs to the effect that one is conforming to a norm. One may be sanctioned for violating a norm, so forming a true belief about whether one has violated a norm (hence potentiating such a sanction) is valuable, regardless of whether one endorses the norm. The chairman in the HELP condition, for example, does not need to say to himself, "Wait! I need to stop and think carefully about whether helping the environment is something that I should be doing." In the HARM condition, however, an inner monologue like this might well be appropriate. The same seems to hold for the CEO who is considering violating or fulfilling a racial identification law in Nazi Germany and indeed for any of the other protagonists in the Knobe effect literature. As Table 1 illustrates, the expected payoffs for violating and conforming to a norm differ.

	Violate norm	**Conform to norm**
True belief about action	good	good
No belief about action	bad	good
False belief about action	bad	bad

Table 1: Expected Payoffs for Norm-Violation and -Conformity × Belief Formation

The middle row is the key. Forming a true belief about one's action enables one to benefit from it (if it conforms to the norm) or at least not be sanctioned for it (if it violates the norm), and forming a false belief about one's action will lead to problems regardless of whether it violates or conforms to the norm. By contrast, it is safe to forget about conforming to a norm but unsafe to forget about violating one. Thus, rational people can be expected to act in accordance with

(NBF) and rational attributors can be expected to attribute psychological attitudes in accordance with (NBA).[3,4]

The central points highlighted by Table 1 and the foregoing discussion of (NBF) and (NBA) generalize. In cases that do not involve the clear violation of any norm, if the expected outcomes of different possible actions are such that not forming a belief about the outcome of one action can lead to significantly different consequences than not forming a belief about another, we claim that the same kind of asymmetries that characterize the Knobe effects will be found. This, we claim, is due to the fact that having practical reasons for paying attention to and considering certain possibilities will generally lead rational agents to form beliefs about those possibilities and in general will lead rational observers to attribute beliefs about those possibilities to agents as well.

We therefore disagree with Knobe's (2010) frequently repeated claim that the contrasting HELP and HARM (or norm-conforming and -violating) conditions in side-effect experiments "are exactly the same in almost every respect but differ in their moral status" (p. 317). On the contrary, we contend that ordinary participants rationally take the psychological processes of protagonists in the contrasting experimental conditions to differ in significant respects and that at the root of these many differences lies a fundamental difference in reflection and belief.[5]

---

[3] Since (NBA) and (NBF) allow for ignoring some evidence, they violate orthodox Bayesian updating and are therefore not purely rational. They may nevertheless be respectable heuristics. After all, orthodox Bayesian updating is costly. It takes a lot of processing power to change one's beliefs in light of every piece of evidence. Mitigating this cost by ignoring pragmatically irrelevant evidence may therefore be a prudent response.

[4] Thus, although we do not endorse the details of Hindriks's (2008) account of the Knobe effect, we do agree with his general point that a proper explanation should take into account not only the mental states (e.g., beliefs, motivations) that the chairman and other protagonists actually have; it should also take into account the mental states that they ought to have.

[5] Thus, we agree with Charles Kalish (2006), who argues, "The studies described by Knobe & Burra, Malle, and Nadelhoffer are taken to demonstrate that the same mental process may be described as acting intentionally or not depending on our evaluation of the outcome. It is not clear, though, that the same causal process is involved in the negative, positive, and neutral stories. In Knobe's (Knobe, 2003[a]) negative outcome scenario the executive considers and rejects a reason not to implement the program. In the positive outcome the executive

3. The Explanatory Power of Belief-Attribution and -Formation Heuristics

Above we noted that Knobe effects have been found with quite the variety of permutations of psychological attitudes and norms. To demonstrate the explanatory power of our account, we show how it applies it to some of the other more prominent cases in the literature.

We believe that any serious contender for explaining the side-effect effects should aim to be comprehensive. Besides the original Knobe (2003a) study, it should be able to explain all of the following studies:

1. Knobe and Mendlow's (2004) **Prudential Norm Violation** study
2. Knobe's (2004a) **Aesthetic Norm Violation** study
3. Knobe's (2007) **Competing Norms** study
4. Nadelhoffer's (2004) **Praiseworthy Protagonist** study
5. Tannenbaum, Ditto, and Pizarro's (2007) **Desire Attribution** study
6. Beebe and Buckwalter's (2010), Beebe and Jensen's (forthcoming) and Beebe's (forthcoming b) **Knowledge Attribution** studies
7. Beebe's (forthcoming a) **Belief Attribution** study
8. Pellizzoni, Girotto, and Surian (2010) **Lack of Asymmetry** studies

This section contains brief descriptions of these studies and how our theory can account for them.

---

recognizes an additional reason to go ahead with the program. Those are two different decision processes." (pp. 197-198)

3.1. Knobe and Mendlow's (2004) ***Prudential Norm Violation*** *study*

In an effort to show that blame is not the primary factor driving the effect, Knobe and Mendlow (2004) presented participants with the following vignette in which the protagonist brings about a potentially bad side-effect but is not blameworthy for doing so:

> Susan is the president of a major computer corporation. One day, her assistant comes to her and says, "We are thinking of implementing a new program. If we actually do implement it, we will be increasing sales in Massachusetts but decreasing sales in New Jersey." Susan thinks, "According to my calculations, the losses we sustain in New Jersey should be a little bit smaller than the gains we make in Massachusetts. I guess the best course of action would be to approve the program." "All right," she says. "Let's implement the program. So we'll be increasing sales in Massachusetts but decreasing sales in New Jersey."

Respondents were asked (a) whether Susan deserves any praise or blame for decreasing sales in New Jersey and (b) whether Susan intentionally decreased sales in New Jersey. Participants replied that Susan deserved neither praise nor blame but that she nevertheless intentionally decreased sales in New Jersey. The results suggest that attributions of praise or blame may not be the fundamental factor underlying the asymmetry in attribution of intentionality. According to our theory, regardless of whether agents engage in praiseworthy or blameworthy behavior, if their actions require greater degrees of deliberation, participants will be more likely to consider them intentional.

3.2. Knobe's (2004a) ***Aesthetic Norm Violation*** *study*

When the chairman makes a decision that will harm the environment, he violates an ethical norm. In order to show that more general evaluative considerations can have similar effects, Knobe (2004a) constructed a vignette in which the goodness or badness of the side-effect was aesthetic rather than moral:

> The Vice-President of a movie studio was talking with the CEO. The Vice-President said: "We are thinking of implementing a new policy. If we implement the policy, it will increase profits for our corporation, but it will also make our movies *better/worse* from an artistic standpoint." The CEO said: "Look, I know that we'll be making the movies *better/worse* from an artistic standpoint, but I don't care one bit about that. All I care about is making as much profit as I can. Let's implement the new policy." They implemented the policy. Sure enough, the policy made the movies *better/worse* from an artistic standpoint.

When participants were asked, "Did the CEO intentionally make the movies *better/worse* from an artistic standpoint?" significantly more participants in the 'worse' condition indicated that the CEO acted intentionally than in the 'better' condition thought so. These results are quite easily explained on our account: a decision that results in aesthetically poorer movies' being made is not something that should be made rashly by a movie studio executive, whereas the same does not seem to be true for a decision that leads to better movies or that leaves movie quality unaffected.

3.3. Knobe's (2007) ***Competing Norms*** *study*

In standard Knobe effect cases, there is only one norm that an agent's action fulfills or violates. However, Knobe (2007) presented participants with the following vignette in which there are two competing norms in each condition, such that both cannot be fulfilled at the same time:

> In Nazi Germany, there was a law called the 'racial identification law.' The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps.
>
> Shortly after this law was passed, the CEO of a small corporation decided to make certain organizational changes.
>
> The Vice-President of the corporation said: "By making those changes, you'll definitely be increasing our profits. But you'll also be *violating/fulfilling* the requirements of the racial identification law."
>
> The CEO said: "Look, I know that I'll be *violating/fulfilling* the requirements of the law, but I don't care one bit about that. All I care about is making as much profit as I can. Let's make those organizational changes!"
>
> As soon as the CEO gave this order, the corporation began making the organizational changes.

Despite the fact that participants found the CEO equally blameworthy in both conditions, respondents were much more willing to attribute intentionality in the 'violate' condition than in the 'fulfill' condition. Given our doxastic heuristics theory, this result shouldn't be surprising. The cost for violating the law will directly affect the CEO, and so we should expect him to reflect on violating more than on fulfilling it.

*3.4. Nadelhoffer's (2004) **Praiseworthy Protagonist** study*

Nadelhoffer (2004) investigated whether a praiseworthy protagonist who violated a prudential norm in order to benefit a friend would receive the same increased attributions of intentionality observed when the protagonist was blameworthy. He used the following vignette to find out:

> Imagine that Steve and Jason are two friends who are competing against one another in an essay competition. Jason decides to help Steve edit his essay. Ellen, a mutual friend, says, "Don't you realize that if you help Steve, you will decrease your own chances of winning the competition?" Jason responds, "I know that helping Steve decreases my chances of winning, but I don't care at all about that. I just want to help my friend!" Sure enough, Steve wins the competition because of Jason's help.

Respondents were asked (a) how much praise Jason deserves and (b) whether Jason *intentionally* decreased his own chances of winning. Nadelhoffer found that participants regarded the side-effect as intentional and considered Jason praiseworthy for his action. This result is different from (Knobe 2003a), which focused only on the asymmetry for blameworthy side-effects. Our theory predicts and explains this result, since deciding to do something that will decrease one's chances of winning requires more careful reflection and will engage one's belief-forming processes more than deciding to do something that will have the opposite effect.

3.5. Tannenbaum, Ditto, and Pizarro's (2007) ***Desire Attribution*** *study*

Tannenbaum, Ditto, and Pizzarro (2007) replicated Knobe's original study with the same HELP and HARM vignettes, but they also added a few additional questions. For instance, they asked respondents, "Did the chairman have a desire to harm/help the environment?" Interestingly enough, their results show that respondents are significantly more willing to attribute a desire to harm the environment than to help the environment. The importance of this study is that we now have an asymmetry for a different sort of psychological property besides intentionality. The more psychological properties that we find exhibiting the side-effect effect, the greater the call is for a theory that unifies all of them. Our doxastic heuristics theory does so because one cannot desire to bring about an effect by performing an action without believing that the action raises the probability of the effect.

3.6. Beebe and Buckwalter's (2010), Beebe and Jensen's (forthcoming) and Beebe's (forthcoming b) ***Knowledge Attribution*** *studies*

Beebe and Buckwalter (2010) also used Knobe's HELP and HARM vignettes, but they asked participants whether the chairman *knew* the new program would help/harm the environment. Again, they found an asymmetry. Respondents were significantly more likely to attribute knowledge to the chairman in the harm case than the help case.

Beebe (forthcoming b) also found high knowledge attributions when the chairman's response was changed from "I don't care at all about harming the environment" to the following:

> I truly wish that I could make money for this company without harming the environment. Unfortunately, that seems to be impossible. Reluctantly, I'm instructing you to start the new program. (adapted from Mele 2006)

The reluctant chairman is clearly less blameworthy than the chairman who doesn't care about harming the environment, but participants think the reluctant chairman is just as likely (if not more likely) to know that his decision will harm the environment. Again, this can be explained in terms of the significant level of reflection that is indicated by his response.

Beebe and Jensen (forthcoming a) extended these results, finding the same kind of asymmetry in knowledge attributions using variants of the case of Susan decreasing sales in New Jersey, the movie studio executive who increased or decreased the quality of his movies from an artistic perspective, and the CEO in Nazi Germany who violated or fulfilled the racial identification law. Knowledge attributions were higher in each of the conditions in which intentionality attributions had been found to be higher. Beebe (forthcoming b) reports similar responses to the following vignette:

> Steve and Jason are two friends who are competing in two different debate competitions. Jason decides to help Steve prepare for his debate. Ellen, a mutual friend, says, "Don't you realize that by spending so much time helping Steve with his debate you are *decreasing/increasing* your chances of winning your own debate?" Jason responds, "I don't care at all about that. I just want to help my friend!" Steve went on to win his debate competition because of Jason's help, *but Jason did not win his/and Jason won his own competition as well.*

When participants were asked whether Jason knew that he was increasing or decreasing his own chances of winning by helping Steve, they were significantly more likely to think that Steve knew he was decreasing his chances than increasing his chances.

As we will explain in detail below, most explanations of the side-effect effect assume that the central protagonists in the experimental vignettes know that their actions will have the good or bad side-effects predicted. However, this common assumption is clearly disproved by the work of Beebe and his collaborators.

We contend that a greater degree of reflection is in general rationally required of agents whose actions will bring about a harm or violate a salient norm, and that this greater degree of

reflection leads attributors to ascribe higher degrees of belief to these agents. Since knowledge entails or at least presupposes belief, this increased willingness to attribute belief should lead to an increased willingness to attribute knowledge, just as the doxastic heuristics theory predicts.

*3.7. Beebe's (forthcoming a) **Belief Attribution** study*

Beebe (forthcoming a) added yet another psychological property to the growing list of properties that are asymmetrically ascribed in Knobe effect cases. Using the original chairman-and-environment case, as well as the cases involving decreased sales in New Jersey, the movie studio executive, and the CEO in Nazi Germany cases, Beebe asked participants whether the protagonists believed or should have believed that the predicted side effects of their actions would come about. The familiar pattern of responses was found yet again, this time providing even more direct support for the doxastic heuristics theory.

*3.8. Pellizzoni, Girotto, and Surian (2010) **Lack of Asymmetry** studies*

A final set of studies corroborating the doxastic heuristics explanation derives from two experiments by Pellizzoni, Girotto, and Surian (2010) that are variations on Knobe (2003a)'s original HELP and HARM cases. In the first, the vice president only mentions the increase in profit while saying nothing about how the new program will help or harm the environment. In the second, the vice president lies to the CEO about the environmental impact. In both studies, the asymmetry in intentionality disappears. This is precisely the result our theory expects. If the

CEO in the HARM case did not believe the environment would be harmed (because he was never told or was lied to), there is no reason to think he harmed the environment intentionally.

4. Doxastic Heuristics Versus Competing Explanations

We contend that any proposed explanation of the side-effect effects must be consistent with the studies described in the previous section. In this section, we show that none of the proposals currently in the literature can explain all of them, whereas our doxastic heuristics theory can. In particular, many prominent explanations of the Knobe effect can be shown to be inadequate because they sharply conflict with the work of Beebe and his collaborators on attributions of knowledge and belief in Knobe effect cases. We contend that our account also reveals how those accounts that are not demonstrably false offer only limited or partial explanations of different aspects of a single, underlying phenomenon.

4.1. Conceptual Competence

Knobe's (2003a, 2003b) original explanation of the side-effect effect was that there was in fact very little to explain. He claimed that the asymmetry he found with intentionality attributions was appropriate because the concept of an intentional action is essentially moralized. In other words, in addition to prescribing that attributors consider an agent's beliefs, desires, intentions or skill when making intentionality judgments, the concept of intentional action also enjoins that they take account of the moral valence of an action's effects. This theory is now widely viewed as having been refuted by a number of experimental results that have come to light since 2003,

and indeed even Knobe (2007) seemed to reject it (though in later papers he seems to have shifted back to his original position). Moreover, the explanation has always met with strong resistance from philosophers who have been unwilling to endorse the idea that moral goodness or badness partially determines the conditions for the correct application of the concept of intentional action. Our explanation of the side-effect effects shows how participants' responses can be viewed as correct, rational, or appropriate without supporting Knobe's controversial claims about the semantics of intentionality attributions.

As we noted above, one of the central ways in which Knobe's original account goes wrong is by assuming that the only relevant difference between the help and harm conditions is the goodness or badness of the side-effect. Indeed, almost every explanation of the side-effect effect on offer assumes that the chairman's state of mind in the help and harm conditions does not differ in any significant respect. For example, Pettit & Knobe (2009) claim:

> Yet it seems that the agent's mental states do not differ between the two cases. The main difference lies instead in the moral status of the side-effect itself. Hence, most researchers have concluded that people's moral judgments are somehow influencing their intuitions as to whether or not an agent acts intentionally. (Pettit & Knobe, 2009, p. 589)

And in his (2010, p. 328), Knobe maintains that "the whole process [of attributing mental states] is suffused with moral considerations from the very beginning." Practically every scholar working on the side-effect effect agrees and has referred to the good and bad (or norm-fulfilling and norm-violating) side-effect actions as "foreseen but undesired" by the agents in question.[6] In other words, these scholars make the untested empirical assumption that there will be no knowledge or belief asymmetry in the relevant cases. The work of Beebe and his colleagues, however, shows that ordinary participants do not take the central protagonists' knowledge or beliefs to be the same in the contrasting conditions.

[6] Cf. Knobe 2004a; Knobe 2006; Knobe & Burra 2006; Leslie, Knobe & Cohen 2006; Doris, Knobe & Woolfolk 2007; Knobe & Doris 2010; Nadelhoffer 2004; Nadelhoffer 2006; Sverdlik 2004; McCann 2005; Mele 2006; Mele & Cushman 2007; Cushman & Mele 2007; Machery 2008; Hindriks 2008.

4.2 Distortion

Another prominent type of explanation of the Knobe effects that also runs afoul of these recent findings is the distortion account of Nadelhoffer (2006).[7] He contends that something has gone seriously wrong with the cognitive processes underlying intentionality attributions because ascribing intentionality to one agent and denying it of another *when the only relevant difference between their situations is the goodness or badness of their actions' side-effects* is semantically, ethically, and even legally problematic. Nadelhoffer argues that when participants have con-attitudes toward a side effect or the agent who produced, these con-attitudes can interfere with the processing of information about what the agent has done. Hence, if people blame or feel negative emotions towards the protagonist, they will be more likely to say the agent brought about the bad result intentionally even though the correct response is otherwise. However, Nadelhoffer's distortion theory cannot explain the prudential and legal manifestations of the side-effect effect, in which the protagonists' actions are not blameworthy. While it may be plausible to think that con-attitudes can bias attributions of intentionality, desire, knowledge, etc., it is less plausible to think that con-attitudes are evoked when someone violates a prudential norm in order to help someone else, or when someone violates a bad law in order to do a morally good act. If we are looking for a fully general account of the side-effect effect, then, the distortion theory will not suffice.

[7] Cf. also Malle (2006; Malle & Nelson, 2003) and Alicke (2008). Malle's attention based version fails to explain the asymmetry at all. He argues that the vignettes used in side-effect studies force participants to focus their attention on the norm-violating behavior, and then, when they are asked about the protagonist's intentionality (or whatever other psychological property), they feel they must somehow use the information to which they have attended. Were their attention not forced in this way, they would not exhibit the asymmetry. However, attentional framing effects occur in both the norm-conformity and norm-violation conditions of the vignettes. So by his account, no asymmetry should be observed.

4.3 Conversational Pragmatics

While the conceptual competence and distortion explanations both assume that participants' responses to the side-effect experiments represent their judgments about intentionality (whether or not those judgments are correct), Adams and Steadman (2004a, 2004b, 2007) think that the responses themselves are suspect, in the sense that they represent not what participants really think but only what they are inclined to say for fear of generating an unintended conversational implicature. Adams and Steadman argue that participants think both that the chairman is blameworthy in the HARM condition and that if they say he did not act intentionally they would be understood to mean that he is not blameworthy. Not wanting to be so interpreted, participants assert (though they perhaps do not genuinely believe) that the protagonist brought about the bad side effect intentionally. Like the distortion theory, the conversational pragmatics theory has a tough time accounting for the side-effect effect in prudential and legal cases. Presumably participants blame the factory-owner who conforms to the racial identification law. Why, then, would they not say that he intentionally complied with the law? Such a statement would generate the implicature that he was to blame. In fact, however, participants exhibit the exact opposite asymmetry: they attribute intentionality when the law is violated (praiseworthy) but not when it is conformed to (blameworthy).

4.4 Semantic Diversity

Knobe's conceptual competence view and the distortion account assume that contrary properties (*bringing about a side-effect intentionally; not bringing about a side-effect intentionally*) are attributed in the good and bad (or norm-conforming and norm-violating) conditions. According to the semantic diversity explanation proposed by Nichols and Ulatowski (2007), by contrast, there are two distinct concepts of intentionality—one having to do with knowledge, the other with motive—that are attributed.[8] Nichols and Ulatowski first presented participants with one of the original Knobe cases and afterward asked them to explain why they answered as they did. Typical answers from participants who said that the chairman intentionally harmed the environment include the following:

"He knew the consequences of his actions before he began."

"He knew that implementing the new program would hurt the environment."

"Because he knew he was going to hurt the environment, but chose to do it anyway."

Typical answers from participants who said that the chairman did not intentionally help the environment include:

"He didn't care. It was an unintended consequence."

"Because his intention was to make money whether or not it will help the environment."

"He didn't INTEND on helping the environment, he INTENDED on making a profit."

Similar explanations were given by those who answered that the chairman either did not intentionally harm the environment or did intentionally help it. Answers in the first group clearly focus on the chairman's knowledge, whereas answers in the second focus on the chairman's motives.

[8] The difference between these two alleged concepts is never spelled out in any detail.

Nichols and Ulatowski assume that participants must believe that the chairman knows the side-effect will occur in both conditions, but it is only when participants employ the concept of intentionality that is based upon knowledge that this recognition of the chairman's knowledge leads them to attribute intentionality. However, because participants do not equally attribute knowledge in both conditions, an important motivation for Nichols and Ulatowski's semantic diversity thesis is undermined. Furthermore, when one considers that Tannenbaum, Ditto, and Pizarro (2007) found the familiar asymmetric pattern of responses in judgments about whether the chairman desired to bring about the side-effect—and indeed when one is reminded of the fact that Knobe effects seem to be found in every type of psychological attribution that is studied—it does not seem at all plausible that participants in fact view the chairman's motives as being exactly the same in the help and harm conditions either.

Finally, while semantic diversity accounts may be consistent with the intentionality attribution asymmetry, they are unable to save all the phenomena. For their theory to apply to all the asymmetries mentioned above, Nichols & Ulatowski would have to propose semantic diversity not just of intentionality but also of knowledge, acting in order to, desire, decision, being in favor of, and advocating (among many others). The wide variety of this list makes the semantic diversity theory *prima facie* implausible. Its inclusion of knowledge makes it a non-starter. Is there one concept of knowledge-as-knowledge and another of knowledge-as-motive? The question stretches the bounds of sense.[9]

[9] Fiery Cushman and Al Mele (2007) also defend a semantic diversity thesis that is unable to handle asymmetries for knowledge and belief.

4.5 Trade-off

According to Edouard Machery's (2008) 'trade-off hypothesis,' participants view protagonists as construing bad side effects as costs that must be incurred in order to receive certain benefits. He claims that participants conceptualize the HARM case as an instance of the chairman being willing to incur a cost in order to get a benefit. Likewise participants supposedly view the HELP case as an instance in which there is no cost/benefit trade-off.

Machery claims further support for his view from the results of an experimental case in which a protagonist must incur a literal financial cost to attain a desired end (the 'extra dollar case').[10] However, recent work by Phelan and Sarkissian (2009), Mallon (forthcoming), and Beebe and Jensen (forthcoming) has shown that participants are strongly disposed to attribute intentionality and knowledge in cases where protagonists clearly do not view the side-effects of their actions as costs. These results, combined with the fact that the trade-off hypothesis is formulated only in terms of attributions of intentionality, mean that the trade-off hypothesis cannot serve as a full and adequate account of all the Knobe effect cases.

4.6 Counterfactual Guidance

One of the most recent contributions to this literature is Holton's (2010) explanation of the side-effect effect in terms of the difference between intentionally following a norm and intentionally violating one. Intentional violation merely requires knowledge that one is doing so. Intentional conformity, on the other hand, requires both knowledge of conforming and being

[10] It is generally recognized that Machery's extra dollar case is not a side-effect effect study, since the extra dollar it cost to receive the benefit is means rather than a side effect. Cf. Mallon (2008) and Cole Wright and Bengson (2009) on this point.

counterfactually guided by the norm, i.e., if the norm were different, the protagonist would conform to it instead of the norm to which he actually conformed. The asymmetry occurs because in the HARM case, the chairman is knowingly violating the norm, while in the HELP case the chairman merely happens to be conform to the norm without being counterfactually guided by it. Hence, only the former affects the environment intentionally.

Like the doxastic heuristics, conceptual competence, pragmatics, and semantic diversity theories (and unlike the distortion theory), the counterfactual-guidance explanation preserves the rationality of participants who exhibit an asymmetry in their attributions. Like the doxastic heuristics account, it also recognizes that norms play an important role in generating the side-effect effects. However, it is ultimately inconsistent with the knowledge and belief side-effect effects. The counterfactual-guidance theory presupposes that people would attribute knowledge and belief equally in the norm-conformity and norm-violation conditions, but as Beebe and his collaborators have demonstrated, they do not.

5. New Experimental Evidence

It's one thing to construct a theory that can explain existing data. It is something else to successfully predict and explain novel data. If our account really is the best explanation of the side-effect effect, it should be able to make such predictions. In this section, we describe some new experiments that provide further confirmation for our theory.

According to the doxastic heuristics account, the practical decision structure of the side-effect effect cases leads participants to attribute greater degrees of reflection and belief to protagonists considering courses of action that violate social norms or have potentially high

practical costs. Our explanation predicts that participants should also be more strongly inclined to think that protagonists remember these facts about norm violations and practical costs at a later time. To test this prediction, 124 undergraduates from a large public university in the northeastern United States were given either the HELP or the HARM version of Knobe's original vignette about the chairman or the 'fulfill' or the 'violate' version of the vignette about the CEO in Nazi Germany. Participants in all four conditions of this between-subjects design were asked, "On a scale of 0 to 10, how likely do you think it is that the chairman remembered that the new program was supposed to *help/harm* the environment?" or "On a scale of 0 to 10, how likely do you think it is that the CEO remembered that the organizational changes were supposed to *fulfill/violate* the racial identification law?" '0' was marked 'Highly Unlikely,' '5' was marked 'Neither Likely nor Unlikely' and '10' 'Highly Likely.'

Just as our theory predicted, participants were significantly more likely to think the chairman remembered harming the environment ($\bar{x}$ = 4.93) than they were to think the chairman remembered helping it ($\bar{x}$ = 2.45).[11] They were also more likely to think the CEO remembered violating the racial identification law ($\bar{x}$ = 5.91) than they were to think he remembered fulfilling the law ($\bar{x}$ = 3.63).[12] This extension of Knobe effect findings from attributions of intentionality, knowledge and belief to attributions of remembering is more easily explained by our account than by competing explanations.

Also according to our hypothesis, vignettes involving conformity to and violation of obviously amoral norms should generate asymmetric attributions of psychological properties. To test this prediction, we conducted two studies, one involving a *conventional* norm, the other

[11] t (40) = 3.488, $p < 0.005$.
[12] t (60) = -3.234, $p < 0.005$.

involving a *descriptive* norm.[13] The participants in the conventional norm case were 60 adult "workers" on the Amazon Mechanical Turk (www.mturk.com), which coordinates online quizzes and surveys. Participants were randomly assigned to read either the conformity or the violation version of the following vignette:

> William is from Scotland, where it is customary for a man to wear a kilt on a special occasion. William is planning to attend a friend's wedding in *Calabria/Edinburgh*, *an Italian/a Scottish* city. William says to his friend David, "I'm thinking of wearing this kilt to the wedding." David responds, "If you do that, you *won't/will* be following Italian/Scottish custom." William says, "I don't care at all about *Italian/Scottish* custom. I just want to wear this kilt." William *does/doesn't* wear the kilt at the wedding.

Participants in the descriptive norm study were assigned to read either the conformity or the violation version of the following vignette:

> Jessica lives in a neighborhood where everyone (including Jessica herself) happens to own a dog. One afternoon, she is planning to go for a walk and decides *not to/to* take her dog. Her friend Aaron says, "Jessica, if you go out like that, you *will/won't* be doing what everyone else is doing." Jessica responds, "I don't care at all what everyone else is doing. I just want to go for a walk *without/with* my dog." She goes ahead with her plan, and sure enough, she ends up doing what *no one/everyone* else is doing.

Participants were then asked to rate their level of agreement on a 7-point Likert scale with the statement, "William intentionally *violated/conformed to Italian/Scottish* custom" (for the conventional case) or "Jessica intentionally did what *no one/everyone* else was doing." As predicted, participants were significantly more likely to think that William intentionally violated Italian custom ($\bar{x}$ = 5.07) than that he intentionally conformed to Scottish custom ($\bar{x}$ = 3.87),[14] and they were significantly more likely to think that Jessica intentionally did what no one else was doing ($\bar{x}$ = 4.90) than that she intentionally did what everyone else was doing ($\bar{x}$ = 3.73).[15] Since the doxastic heuristics theory predicts asymmetries across the board whenever the protagonist performs an action that requires extra deliberation or reflection, we claim further confirmation from these results. Violating a social norm (even a norm that merely involves

[13] For more on the distinction between conventional and descriptive norms, see Bicchieri (2006).
[14] $t(60) = 2.053$, $p < 0.05$)
[15] $t(118) = 2.478$, $p < 0.05$)

fashion) rationally requires more deliberation than conforming to one. Doing what everyone else is doing rationally requires less reflection than doing what no one else is doing.

6. The promise of unification

Because of the manifest failure to date of many 'silver bullet' or single-factor attempts at explaining the side-effect effects, some scholars have grown skeptical of there being any simple, elegant and unifying account that has sufficient explanatory breadth and power. Mark Phelan and Hagop Sarkissian (2009), for example, suggest that "given the number of variables at play, any parsimonious account of the relevant data is implausible," and after complaining about the "over-simplicity of existing accounts" that focus on only one or two features of the situations depicted in Knobe effect cases, they conclude:

> As the debate over intentional side effects stands, though, we must conclude that attempts to account for the Knobe effect by recourse to only one or two variables, though instructive, are incomplete and overreaching in their ambition. It is time to abandon the dream of parsimony. (pp. 164, 179)

We have endeavored to keep the dream of parsimony alive by offering a simple, straightforward, and unified account of the data in Knobe effect cases that involve not only intentional action, but also desire, knowledge, belief and remembering. We contend that the asymmetry for belief exists for good reason: violating a norm typically comes with costs (real or potential), while conforming typically does not. So you should have accurate beliefs about what might happen when you violate a norm, but not necessarily when you conform. This asymmetry for belief is the underlying cause of all other observed asymmetries because intention, desire, knowledge and remembering all presuppose belief.

A prominent and perennial theme in Knobe's own writings on the side-effect effect is that recent experimental results have shown that the most widely accepted view about the nature of folk psychology is fundamentally flawed. For example, he writes:

> There is something extremely plausible and convincing about the claim that folk psychology should be seen as a tool for predicting, explaining and controlling behavior. Nonetheless, I think we now have good reason to believe that this claim is not quite right. As I shall try to show here, certain aspects of folk psychology appear to have been shaped in a very fundamental way by other, very different uses. (2006, p. 204)

> To some degree at least, it seems that these results should come as a surprise to those who think of people's concept of intentional action as a tool for predicting, controlling and explaining behavior. After all, it seems that the best way to accomplish these 'scientific' goals would be to ignore all the moral issues and focus entirely on a different sort of question (e.g., on questions about the agent's mental states). How then are we to make sense of the fact that moral considerations sometimes influence people's application of the concept of intentional action? (2006, p. 207)

> Law courts commonly assume that judgments of purpose are purely factual judgments to be decided by juries. The same assumption is made in the literature on the child's theory of mind. The side-effect effect suggests, however, that such judgments may sometimes be partly factual and partly moral. To the extent that such judgments are moral, theory of mind is unlike a scientific theory, and its development is not reducible to discovering matters of fact. (Leslie, Knobe & Cohen, 2006, p. 426)

Knobe (2010) suggests that on the traditional view of folk psychology, "people truly are engaged in an effort to pursue something like a scientific investigation, but that they simply aren't doing a very good job of it" because, although "the competencies underlying people's judgments actually are purely scientific in nature […] there are then various additional factors that get in the way of people's ability to apply these competencies correctly." (p. 315)

Behind each of the claims is Knobe's view that the psychological attitudes of protagonists in contrasting experimental conditions are not (and are not taken to be) significantly different:

> It might at first appear that people's use of this distinction [between intentional and unintentional action] depends entirely on certain facts about the role of the agent's mental states in his or her behavior, but experimental studies consistently indicate that something more complex is actually at work here. It seems that people's moral judgments can somehow influence their intuitions about whether a behavior is intentional or unintentional. (2010, p. 317)

According to our account, however, people are indeed basing their attributions of intentional and unintentional action on "certain facts about the role of the agent's mental states in his or her

behavior"—in particular, mental states of knowledge and belief that Beebe and his collaborators have shown are not equally ascribed to protagonists in contrasting conditions.

We thus agree with Kevin Uttich and Tania Lombrozo (2010), who claim:

> Our approach concedes that moral judgments influence ToM [i.e., theory of mind], 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. (pp. 88-89)

Accordingly, we do not think the Knobe effect requires us to rethink the widely shared view that human folk psychology has been shaped by its fundamental role in the prediction, explanation and control of other's behavior. In fact, our view seems to be compatible with and complementary to that of Uttich and Lombrozo, who offer an explanation based on the fact that behavior that conforms to norms is generally less informative about the mental states underlying the behavior than behavior that violates norms. Because norms automatically provide reasons for acting in accord with them, Uttich and Lombrozo argue that norm-violating behavior requires and often points toward reasons for the norm violation. We agree and suggest that differences in an action's practical costs and benefits provide important information about what those reasons are likely to be.

References

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

Adams, F., & Steadman, A. (2004b). Intentional action and moral considerations: Still pragmatic. *Analysis*, 64, 268-276.

Adams, F., & Steadman, A. (2007). Folk concepts, surveys, and intentional action. In C. Lumer (Ed.), *Intentionality, deliberation, and autonomy: The action-theoretic basis of practical philosophy* (pp. 17-33). Aldershot: Ashgate.

Alicke, M.D. (2008). Blaming badly. *Journal of Cognition and Culture*, 8: 179–186.

Beebe, J. R. (forthcoming a). A Knobe effect for belief ascriptions.

Beebe, J. R. (forthcoming b). Attributions of knowledge and blame in Knobe effect cases.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language*, 25, 474-498.

Beebe, J. R., & Jensen, M. (forthcoming). Surprising connections between knowledge and action: The robustness of the epistemic side-effect effect. *Philosophical Psychology*.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms.* Cambridge: Cambridge University Press.

Cushman, F., & Mele, A. (2007). "Intentional action: Two-and-a-half folk concepts?" In J. Knobe and S. Nichols (Eds.), *Experimental Philosophy* (pp. 171-188). New York: Oxford University Press.

Doris, J., Knobe, J., & Woolfolk, R. L. (2007). Variantism about responsibility. *Philosophical Perspectives*, 21, 183-214.

Hindriks, F. (2008). Intentional action and the praise-blame asymmetry. *Philosophical Quarterly*, 58, 630-641.

Holton, R. (2010). Norms and the Knobe effect. *Analysis*, 70, 417-424.

Kalish, C. W. (2006). Integrating normative and psychological knowledge: What should we be thinking about? *Journal of Cognition and Culture*, 6, 191-208.

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, 309-24.

Knobe, J. (2004a). Folk psychology and folk morality: Response to critics. *Journal of Theoretical and Philosophical Psychology,* 24, 270-279.

Knobe, J. (2004b). 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, 315-329.

Knobe, J. & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture*, 6, 113-132.

Knobe, J., & Mendlow, G. (2004). The good, the bad and the blameworthy: Understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology*, 24, 252-258.

Knobe, J., & Doris, J. (forthcoming). Strawsonian variations: Folk morality and the search for a unified theory. In J. Doris and the Moral Psychology Research Group (Eds.), *The Handbook of Moral Psychology.* Oxford: Oxford University Press.

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science,* 17, 421-427.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language*, 23, 165-189.

Malle, B. F. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture,* 6, 87-113.

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. (forthcoming). Knobe vs. Machery: Testing the trade-off hypothesis. *Mind & Language*.

McCann, H. (2005). Intentional action and intending: Recent empirical studies. *Philosophical Psychology*, 18, 737–748.

Mele, A. (2006). The folk concept of intentional action: A commentary. *Journal of Cognition and Culture,* 6, 277-290.

Mele, A., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, 31: 184-201.

Nadelhoffer, T. (2004a). On praise, side effects, and folk ascriptions of intentionality. *Journal of Theoretical and Philosophical Psychology,* 24, 196-213.

Nadelhoffer, T. (2004b). Blame, badness, and intentional action: A reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology,* 24, 259-269.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality. *Philosophical Explorations*, 9, 203-220.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind and Language*, 22, 346-365.

Pellizzoni, S., Girotto, V., & Surian, L. (2010). Beliefs and moral valence affect intentionality attributions: The case of side effects. *Review of Philosophy and Psychology*, 1, 201-209.

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

Phelan, M., & Sarkissian, H. (2008). 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.

Phelan, M., & Sarkissian, H. (2009). Is the 'trade-off hypothesis' worth trading for? *Mind & Language*, 24, 164-180.

Sverdlik, S. (2004). Intentionality and moral judgments in commonsense thought about action. *Journal of Theoretical and Philosophical Psychology*, 24, 224-236.

Tannenbaum, D., Ditto, P.H. & Pizarro, D.A. (2007). Different moral values produce different judgments of intentional action. Unpublished manuscript, University of California-Irvine.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116, 87-100.