Personality and Social
Psychology Bulletin
36(12) 1635–1647

DOI: 10.1177/0146167210386733
http://pspb.sagepub.com

Can Unintended Side Effects Be Intentional? Resolving a Controversy Over Intentionality and Morality

Steve Guglielmo[1] and Bertram F. Malle[1]

Abstract

Can an event's blameworthiness distort whether people see it as intentional? In controversial recent studies, people judged a behavior's negative side effect intentional even though the agent allegedly had no desire for it to occur. Such a judgment contradicts the standard assumption that desire is a necessary condition of intentionality, and it raises concerns about assessments of intentionality in legal settings. Six studies examined whether blameworthy events distort intentionality judgments. Studies 1 through 4 show that, counter to recent claims, intentionality judgments are systematically guided by variations in the agent's desire, for moral and nonmoral actions alike. Studies 5 and 6 show that a behavior's negative side effects are rarely seen as intentional once people are allowed to choose from multiple descriptions of the behavior. Specifically, people distinguish between "knowingly" and "intentionally" bringing about a side effect, even for immoral actions. These studies suggest that intentionality judgments are unaffected by a behavior's blameworthiness.

Keywords

intentionality, blame, moral judgment, morality, praise, bias

Received July 21, 2009; revision accepted July 3, 2010

Judging the intentionality of behavior is an important process in social cognition. For one, this process guides people's interpretation of behavior, favoring either reasons or causes to explain the behavior (Malle, 1999). As important, intentionality judgments guide people's moral judgments. In particular, an agent receives substantially more blame for a negative behavior that was intentional than for a similar one that was unintentional (Heider, 1958; Ohtsubo, 2007; Shultz & Wells, 1985). An agent also receives more blame for an unintentional negative outcome that could have been prevented (if only the agent had intended to do so) than for a similar outcome that could not have been prevented (Abbey, 1987; Davis, Lehman, Silver, Wortman, & Ellard, 1996; Shaver, 1985; Weiner, 1995). In fact, just as intentionality judgments incorporate information about the agent's beliefs, desires, awareness, and skill (Malle & Knobe, 1997, 2001), so do judgments of blame (Fincham & Jaspars, 1979; Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002). Thus, judgments of intentionality and its components critically inform and constrain judgments of blame (Guglielmo, Monroe, & Malle, 2009; Solan, 2003), suggesting a schematic model of *intentionality → blame.*[1]

This influence of intentionality on blame involves a complex conceptual framework, as intentionality judgments rely on multiple necessary conditions (Kashima, McKintyre, & Clifford, 1998; Malle & Knobe, 1997; Mele, 1992). Malle and Knobe (1997) showed that people deem a behavior intentional only when five distinct components are present (see Figure 1): the agent's *desire* for an outcome, *beliefs* about the action leading to the outcome, the *intention* to perform the action, *awareness* of the act while performing it, and a sufficient degree of *skill* to reliably perform the action. In several studies, Malle and Knobe (1997) documented that when any one of the five components was absent, people very rarely judged the behavior intentional. Additional research showed that people make systematic distinctions even between such closely connected components as desires and intentions (Malle & Knobe, 2001).

The Challenge

Recent findings, however, question whether people are consistently sensitive to all these intentionality components and whether the standard *intentionality → blame* model is correct.

[1]Brown University, Providence, RI, USA

Corresponding Author:
Steve Guglielmo, 89 Waterman St., Box 1821, Providence, RI 02912
Email: steve_guglielmo@brown.edu

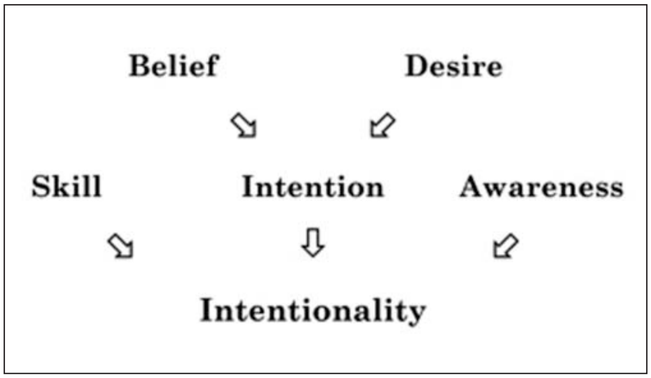

Figure 1. A model of the folk concept of intentionality
From B. F. Malle & J. Knobe. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33,* 101-121. © Lawrence Erlbaum Associates

In particular, Knobe (2003a) claimed that people consider negative side effects (i.e., outcomes that were foreseen but not intended) to be intentional but neutral or positive side effects to be unintentional. To demonstrate this tendency, Knobe contrasted two conditions (one of harming, one of helping) of the following scenario:

> 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 [and] it will also harm [help] the environment." The chairman of the board answered, "I don't care at all about harming [helping] 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 [helped].

Although the chairman knew that the environment would be harmed [helped], people recognize that he did not specifically intend to bring about this effect (Knobe, 2004b; McCann, 2005). Nonetheless, 82% of people said the chairman intentionally harmed the environment whereas only 23% said he intentionally helped the environment. These striking results have been replicated numerous times with identical and different vignettes (Knobe, 2004b; Mallon, 2008; Nadelhoffer, 2004, 2005; Nichols & Ulatowski, 2007; Wright & Bengson, 2009). We refer to this body of work as the "side-effect findings."

In Knobe's (2003a, 2005) and others' interpretation (Alicke, 2008; Cokely & Feltz, 2009; Nadelhoffer, 2006a; Wright & Bengson, 2009), people's intentionality judgments about negative side effects are biased—specifically, the blameworthiness of the harming chairman's behavior caused people to see the unintended side effect as intentional. Indeed, blame judgments (in the harm condition) were more extreme than praise judgments (in the help condition), and they predicted intentionality judgments. Knobe (2003a, 2005) therefore proposed a reverse model: *blame → intentionality.*[2] In other words, people initially assign blame to the agent of a negative behavior, which then informs or biases their judgment about whether the behavior was intentional.

The claim that moral judgments may affect intentionality judgments is hotly debated in the current philosophical and cognitive science literature (e.g., Machery, 2008; Mallon, 2008; Nadelhoffer, 2006a; Phelan & Sarkissian, 2008; Uttich & Lombrozo, 2010; Wright & Bengson, 2009), and this debate is beginning to engage social, personality, and developmental psychology as well (Alicke, 2008; Cokely & Feltz, 2009; Leslie, Knobe, & Cohen, 2006; Malle, 2006). The new *blame → intentionality* model finds kinship among social-psychological theories that emphasize the influence of blame (or immorality) on other cognitive processes and judgments (Alicke, 2000; Haidt, 2001). Pizarro, Laney, Morris, and Loftus (2006) found that blame can bias people's memory of an event, and Alicke (2000) suggested that negative moral sentiments can lead people to interpret evidence in a biased manner. For example, Alicke (1992) showed that a person who was speeding to hide a vial of cocaine was judged to have more responsibility for his ensuing car accident than a person who was speeding to hide his parents' anniversary gift.

Although Knobe's (2003a, 2010) challenge is generally consistent with Alicke's (2000) analysis, it makes a stronger claim. Alicke's model predicts biases in blame, responsibility, and causality judgments, but the pertinent studies do not directly assess whether people unduly consider a negative behavior *intentional*. Knobe (2003a, 2003b, 2010) argued that the concept of intentionality itself is imbued with moral meaning and that intentionality judgments directly reflect the moral goodness or badness of a behavior: "People's intuitions as to whether or not a behavior was performed intentionally can be influenced by their beliefs about the moral status of the behavior itself" (Knobe, 2004a, p. 270).

If correct, the reverse *blame → intentionality* relation would have serious implications for psychological theory. Previous research supporting the standard model of *intentionality → blame* may be the exception rather than the rule, and the five-component model of intentionality (Malle & Knobe, 1997) would be incorrect for negative outcomes. In addition, the integrity of legal decision making would come into question, as juries, which must determine whether a serious criminal behavior was performed intentionally, might be biased by early moral judgments about the behavior in question (e.g., Nadelhoffer, 2006b).

Two Puzzles

In this article, we try to resolve two related puzzles entailed by the side-effect findings. The first puzzle is why people's intentionality judgments differ across the harming and helping conditions. To solve this puzzle, we challenge the assumption that the negative and positive scenarios in the extant findings differ only in their moral valence. Study 1 shows that the harming and the helping conditions differ in the protagonist's attitude toward the side effect. Even though the chairman says in both cases, "I don't care at all about . . .," people see him as wanting to harm the environment more than they see him as wanting to help the environment. Study 2 shows that this pattern generalizes to a host of other behaviors. Study 3 reveals a complete side-effect asymmetry identical to Knobe's (2003a) original but featuring a nonmoral action, and the asymmetry is explained by desire inferences. Studies 4a and 4b show that manipulating desire elicits corresponding changes in intentionality judgments: Weakening desire in the harm case reduces intentionality judgments; strengthening desire in the help case increases them.

The second puzzle is why people judge a negative side effect intentional despite viewing it as unintended. We hypothesize that people deem the side effect intentional only when forced into a dichotomous judgment of "intentional" versus "not intentional." Studies 5 and 6 examine whether people still describe the side effect as intentional once they are given multiple descriptions to choose from. Our results show that, in this case, they hardly ever judge unintended (negative or positive) side effects as intentional.

Study 1

One clear difference between Knobe's (2003a) contrasting scenarios is the outcome itself—harm versus help to the environment. But another may be the degree of "desire" or "pro-attitude" (Davidson, 1963) the agent had toward the outcome. People generally expect others to prevent negative and promote positive outcomes. Both chairmen in Knobe's scenarios violate these expectations, but in different ways. By proclaiming "I don't care at all about harming the environment . . .," the harming chairman dismisses any concern about harming the environment and displays approval of the harm, thus a modest desire. In contrast, the helping chairman, although uttering the same words ("I don't care at all about . . ."), does not even welcome any benefit to the environment, thus clearly lacking desire. If people infer different degrees of the agent's desire for the harming versus helping side effects, such a difference would help explain Knobe's original findings. Importantly, this explanation would invoke a central component of people's intentionality concept rather than the impact of moral valence. Study 1 therefore tested the role of desire in side-effect findings.

Method

Participants were 61 undergraduate students who completed a one-page questionnaire as part of a larger computer-presented survey and received partial course credit. Each read either the harming or helping vignette from Knobe's (2003a) study, then answered three questions: *intentionality* ("Did the CEO intentionally harm [help] the environment?"), with possible responses "yes" and "no"; *blame/praise* ("How much blame [praise] does the CEO deserve?"); and *desire* ("To what extent did the CEO want to harm [help] the environment?"), the latter two questions rated on a 0 to 6 scale.[3]

Results and Discussion

The pattern of intentionality responses replicated Knobe's (2003a) findings—87% deemed the harming intentional whereas only 20% deemed the helping intentional, $\chi^2 = 27.6$, $p < .001$. However, desire ratings differed considerably between conditions. Whereas people inferred barely any desire in the helping condition ($M = 1.50$, $SD = 1.33$), they inferred substantially more in the harming condition ($M = 3.55$, $SD = 1.61$), $t(59) = 5.41$, $p < .001$, $d = 1.38$. Moreover, desire ratings were highly correlated with intentionality judgments, $r = .54$—the more that people thought the CEO wanted the outcome, the more they judged it to be intentional.

Blame ($M = 4.97$, $SD = .95$) was moderately correlated with intentionality, $r = .40$, $p < .05$, whereas praise ($M = 2.33$, $SD = 1.47$) was not, $r = .12$, $p > .10$.

Thus, Knobe's (2003a) positive and negative side-effect scenarios differed in one important respect (aside from valence). The agent's professed indifference ("I don't care at all about . . .") was interpreted as moderate desire for the outcome in the negative case but as virtually no desire in the positive case. Furthermore, variation in desire was highly predictive of people's intentionality judgments, consistent with the component model of intentionality (Malle & Knobe, 1997).

A possible alternative account of the results of Study 1 is that people's desire ratings, like their intentionality judgments, may themselves reflect an influence of moral judgment (e.g., Knobe, 2010; Pettit & Knobe, 2009). Both methodological and empirical considerations weaken this possibility. The chairman's declaration of desire ("I don't care at all . . .") precedes both his decision to adopt the program and the outcome itself, making it likely that people arrive at their desire inferences before any moral considerations. In addition, we compared the standard and alternative accounts in two separate structural equation models of the harm data. The standard model predicts an indirect path of desire → intentionality → blame, which was significant, $t(30) = 1.73$ ($p < .05$, one-tailed). In contrast, the alternative model predicts an indirect path of blame → desire → intentionality, which was not significant, $t(30) < 1$. It appears, therefore, that desire is not guided by blame; rather, it enhances blame by way of an intentionality judgment.

We conducted Study 2 to further test our claim that not caring about negative versus positive outcomes conveys different degrees of desire. Even isolated statements of the form "I don't care at all about X" should foster greater desire ratings when X is something negative than when it is something positive, whether of moral significance or not.

Study 2

Method

Participants were 37 undergraduate students who completed a computerized task as part of a larger study on moral cognition, in return for course credit. Participants provided ratings of desire for eight statements. For each one, participants were asked: "Somebody says: 'I don't care at all about [action].' What is the person's likely attitude toward [action]?" They provided their desire ratings on a scale from 1 (*doesn't want to [action]*) to 9 (*wants to [action]*). The statements were presented in one of two fixed quasi-random orders. The four negative actions were: harming the environment, cheating on the exam, burping at the dinner table, and killing wasps. The four positive actions were: helping the environment, studying for the exam, funding education, and giving to charity. Several actions were designed to be morally irrelevant (e.g., burping at the dinner table, studying for the exam), countering the possibility that any observed desire differences can be attributed to moral concerns.

Results and Discussion

As predicted, people inferred substantially greater desire for negative actions (M = 6.02, SD = 1.45) than for positive actions (M = 2.27, SD = 1.15), F(1, 36) = 130.2, $p < .001$, d = 2.87. The mean desire ratings for the individual statements are shown in Table 1.

Thus, "not caring at all" about an action conveys a very different attitude when paired with a negative (norm-violating) action than when paired with a positive (norm-conforming) action. This valence difference, however, does not depend on moral considerations; nonmoral actions showed exactly the same pattern.

Study 3 tested directly whether norm-violating actions paired with an "I don't care" statement are judged more intentional than norm-conforming actions paired with such a statement. An agent either conformed to or broke a trivial social norm (following a dress code), and we expected that people would view breaking the dress code as more intentional than conforming to the dress code. Moreover, this pattern should be explained by differences in inferred desire.

Study 3

Method

Participants were 56 adults who completed the study while waiting at a public transit center. They received no compensation. All participants read both the breaking and the conforming dress code scenarios, whose order was counterbalanced across participants. The scenario read as follows:

> The woman's husband came to her before the party and said: "I have found a dress for you to wear at the party. You will look great, but [and] you will also break [conform to] the party dress code."
>
> The woman answered, "I don't care at all about breaking [conforming to] the party dress code. I just want to look great." She wore the dress and, sure enough, she broke [conformed to] the party dress code.

After each scenario, participants answered a desire question ("How much did the woman want to break [conform to] the party dress code?") on a scale from –5 (*not at all*) to 5 (*very much*), and a yes–no intentionality question ("Did the woman intentionally break [conform to] the party dress code?"). They also answered a question about the importance of dress codes ("How important do you think it is for people to follow party dress codes?") on the same scale as the desire question.

Results and Discussion

As predicted, people inferred greater desire in the norm breaking scenario (M = 1.00) than in the conforming scenario (M = –1.25), F(1, 53) = 16.31, $p < .001$, d = .65. This effect was not moderated by the order of the scenarios, $F < 1.0$.

Intentionality judgments also showed the expected pattern: People were far more likely to say that the agent intentionally broke the dress code (64%) than that she intentionally conformed to it (35%), Wilcoxon z = 3.20, $p < .001$, d = .43. Moreover, there was a strong connection between desire and intentionality judgments, r = .46. In fact, desire mediated the effect of action on intentionality: There was a substantial action → desire → intentionality indirect effect, t(110) = 2.70, $p < .01$, reducing the action → intentionality direct effect to marginal significance, $p < .10$.

Table 1. Desire Ratings for Negative and Positive Actions (Study 2)

	M	(*SD*)
Negative actions		
"I don't care at all about . . ."		
Cheating on the exam	6.83	(1.96)
Burping at the dinner table	6.53	(1.81)
Harming the environment	5.60	(1.44)
Killing wasps	5.22	(2.20)
Positive actions		
"I don't care at all about . . ."		
Funding education	2.87	(1.96)
Giving to charity	2.08	(1.26)
Studying for the exam	2.06	(1.33)
Helping the environment	2.00	(1.43)

Note: All ratings were made on a scale from 1 (*doesn't want to [action]*) to 9 (*wants to [action]*).

Ratings of the importance of following dress codes were fairly low (M = 1.27) and were entirely unrelated to the intentionality of breaking the dress code, r = .06. So even if a small number of participants might have considered a dress code so important as to be almost moral, this perception was unrelated to the major results.

Study 3 showed that people infer different levels of desire for a side effect depending on whether the side effect entails conforming to or breaking a nonmoral social norm. These desire inferences, in turn, are strong predictors of whether the agent brought about the side effect intentionally. These results are consistent with other recent findings of nonmoral intentionality asymmetries (Machery, 2008; Uttich & Lombrozo, 2010), but they extend the literature by highlighting the critical role of desire inferences in attaining these effects.

If Knobe's (2003a) original negative and positive side-effect scenarios elicited different inferences of the agent's desire, and if this difference contributed to the surprisingly frequent intentionality judgments in the negative condition, then two predictions follow. First, weakening the evidence for such a desire should reduce intentionality judgments for negative side effects; Study 4a tested this prediction. Second, strengthening the evidence for the agent's desire in the positive condition should increase intentionality judgments for positive side effects; Study 4b tested this prediction.

Study 4a

Method

Participants were 82 undergraduate students who completed a one-page questionnaire as part of a larger computer-presented survey, in return for partial course credit. They read a modified version of Knobe's (2003a) original harm scenario. Here, a "regretful" CEO said "It would be unfortunate if the environment got harmed. But my primary concern is to increase profits. Let's start the new program." Participants answered the standard yes–no intentionality question ("Did the CEO intentionally harm the environment?").

Results and Discussion

In this new regretful CEO condition, only 40% deemed the harming intentional. This rate is significantly lower than the rates in Knobe's (2003a) original harm condition (82%) and in our Study 1 sample (87%), both χ^2s > 17.0, ps < .01.

We gave a separate sample (N = 50) the same "regretful" CEO story and, in addition to the standard intentionality question, we also asked the desire question from Study 1 ("To what extent did the CEO want to harm the environment?") and the blame question. As expected, desire ratings were indeed lower in the regretful condition (M = 2.14, SD = 1.51) than in the original condition from Study 1 (M = 3.55, SD = 1.61), t(78) = 3.88, p < .001, d = .91. Importantly, intentionality judgments were again lower (59%) than those in Knobe's (2003a) original condition and in our Study 1, both χ^2s > 5.4, ps < .05, and a logistic regression showed that desire ratings predicted intentionality judgments, z = 2.01, p < .05. Blame ratings were as high in the regretful condition (M = 4.88, SD = 1.29) as in the original condition (M = 4.97, SD = .95), p > .10.

Thus, intentionality judgments about negative side effects drop considerably when there is evidence that the agent does not desire the side effect. These results are consistent with other recent findings in the literature: Fewer than 30% of people judged a negative side effect intentional when the protagonist said, "I feel terrible about [side effect]" (Phelan & Sarkissian, 2009) or "I'll definitely regret [side effect]" (Mele & Cushman, 2007). However, in these studies the protagonist had more laudable goals (reducing pollution or fixing a mosquito problem, rather than making profit), thus making it unclear what caused the drop in intentionality. Our study held the (less popular) goal of increasing profits constant and still showed that manipulating desire had a substantial effect on intentionality.[4]

If desire guides people's judgments of intentionality, then just as weakening desire dampens intentionality perceptions in Knobe's (2003a) harm case, so too should strengthening desire enhance intentionality perceptions in the help case. We tested this prediction in Study 4b.

Study 4b

Method

Participants were 39 undergraduate students who volunteered to complete the study while spending time in the campus mailroom. All participants read a modified version of Knobe's original help scenario. Rather than dismissing the ensuing benefit of the program ("I don't care at all about . . ."), the "welcoming" chairman stated, "I'm thrilled about helping the environment! And it's crucial that we increase profits. Let's start the new program." Participants answered the same intentionality, desire, and praise questions as in Study 1 (followed by several new questions, which will be discussed in Study 5).

Results and Discussion

In the welcoming condition, 56% of people judged the helping to be intentional. This rate is substantially higher than that in the original, uncaring help condition (20% from Study 1), χ^2 = 9.32, p < .01. Desire ratings were also higher in the welcoming condition (M = 3.67, SD = 1.15) than in the original condition (M = 1.50, SD = 1.33), t(67) = 7.12, p < .001, and the connection between desire and intentionality was strong, r = .51. Lastly, praise ratings were higher (M = 3.44, SD = 1.05) than in the original condition (M = 2.33, SD = 1.47), t(67) = 3.51, p < .001.

Thus, increasing the agent's desire for the positive side effect boosted people's intentionality judgments just as decreasing the agent's desire for the negative side effect diminished their intentionality judgments. As a result, intentionality rates in the two conditions converged at around 50%.

Interim Conclusion

Studies 1 to 4 challenge the side-effect findings by proposing that they are best explained not by differences in moral valence but by differences in desire. Indeed, the original harm and help scenarios elicited different inferences about the agent's desire for the respective side effects, which in turn strongly predicted intentionality judgments (Study 1). This is because "not caring" about negative outcomes indicates a moderate desire for those outcomes whereas "not caring" about positive outcomes indicates an utter lack of desire for those outcomes (Study 2). This pattern is not limited to the moral domain, as violating even a trivial social norm is judged more intentional than conforming to that norm, and this effect on intentionality is mediated by inferences about the agent's desire (Study 3). Studies 4a and 4b manipulated the agent's desire in Knobe's (2003a) original scenarios and showed expected effects on intentionality. Intentionality judgments dropped to 40-59% in the negative case when the harming CEO regretted the negative side effect, and they increased to 56% in the positive case when the helping CEO welcomed the positive side effect. The two desire-adjusted cases effectively meet at comparable rates of intentionality.

We have therefore explained the first puzzle—that people judge negative side effects as more intentional than positive side effects—by demonstrating the critical role of desire inferences in intentionality judgments. Previous demonstrations of the side-effect asymmetry confounded moral valence with desire strength. When controlling for desire strength, the effect of moral valence all but disappears.

But we still have another puzzle to explain. Most people in Knobe's original harm case said that the chairman intentionally harmed the environment while maintaining that he did not intend to harm it (Knobe, 2004b; McCann, 2005). According to standard models of intentionality, people who judge that an agent brought about an event *intentionally* should also judge that this agent *intended* to bring the event about (Adams, 1986; Malle & Knobe, 1997; Searle, 1983). The side-effect findings seem to disprove this prediction and suggest instead that people perceive morally objectionable side effects as unintended but intentional.

We propose that there is an alternative interpretation of this puzzle. The chairman did several things intentionally: He considered a new program, dismissed its potentially harmful side effects, and adopted the program to raise profits. His relation to the environmental harm is therefore not accidental, and when given a forced-choice question ("Did he do it intentionally or unintentionally?"), the unintentional option is simply not correct, making the intentional option the only acceptable response. But perhaps people conceptualize the agent's behavior as neither intentional nor unintentional; they may feel he harmed the environment *willingly* or *knowingly*, or that he *recklessly allowed* harm to occur.

Adams and Steadman (2007) proposed a similar hypothesis, suggesting that people would rather describe the chairman's behavior as harming the environment "knowingly, *but not* intentionally" than "knowingly and intentionally." However, even when given these two options, 80% of participants chose the second one. We believe that Adams and Steadman rightly highlighted the need for alternative response options to examine how people themselves conceptualize the scenarios. However, in their empirical test, Adams and Steadman still provided only a single dichotomous choice, and one of the two options explicitly negated intentionality, which people may flatly reject.

Study 5 therefore investigated how people conceptualize negative side effects when given a chance to select from among several descriptions. Knobe (2003a) and others suggested that people truly view unintended side effects as intentional, so they should continue to endorse this description when it is presented alongside alternative descriptions. In contrast, the standard intentionality model predicts that people do not view unintended side effects as genuinely intentional, so they should prefer alternative descriptions if given an option to choose them.

Study 5

Method

Participants were 236 undergraduate students who completed a short questionnaire in the classroom or in the campus mailroom. One group (N = 153) read the original harm vignette used in Knobe (2003a) or an indistinguishable variant.[5] Participants then answered the standard yes–no intentionality question, completed ratings of desire, blame/praise, and answered a yes–no intention question ("Did the chairman intend to harm [help] the environment?"). Next, participants were asked: "Which of the following descriptions of the chairman's behavior are correct? For any description that seems correct to you, put a checkmark in the box to the left of it." The five descriptions are listed in Table 2. Finally, participants indicated which of the five statements was the "most accurate" and which was the "second-most accurate" description of the chairman's behavior.

We added a second group (N = 83) who read the modified help vignette from Study 4b, which contained evidence of the agent's desire for the side effect. (The sample of N = 39 from Study 4b is included in this group.) In contrast to the standard help vignette, which does not elicit intentionality judgments (due to a lack of the agent's desire), the "welcoming" help version elicits enough intentionality endorsements (from at least 50% of people) to ask the same question as in

Table 2. Endorsement Percentages by Condition in the Full Sample and Among Those Who Viewed the Side Effect as Unintended (Study 5)

	Full sample		Side effect unintended	
Description	Harm	Help	Harm	Help
"The chairman . . .				
[1] intentionally harmed/helped the environment."	46	45	26	29
[2] intentionally put profits before the environment."	93	29	94	40
[3] intentionally adopted a program he knew would harm/help the environment."	82	77	76	73
[4] intentionally disregarded the environment when adopting the program."	69	1	73	2
[5] intentionally started an environment-harming/helping program."	42	48	31	31

the harm case: Do the forced-choice intentionality judgments persist, and are they thus validated, in the presence of alternative event descriptions?

Results

Harm condition. As expected, when given the yes–no intentionality question, most people judged the harming to be intentional (71%). According to the standard side-effect findings, a strong majority of participants should likewise endorse statement [1] ("The chairman intentionally harmed the environment"). However, only 46% of them did in the "any correct" responses (see Table 2). Similarly, statement [5] ("The chairman intentionally started an environment-harming program"), a variant of [1], was rarely judged correct (42%). In contrast, twice as many people judged statements [2] ("The chairman intentionally put profits before the environment") and [3] ("The chairman intentionally adopted a program he knew would harm the environment") to be correct (93% and 82%, respectively), making each far more popular than [1], both Wilcoxon zs > 6.5, ps < .001.

Previous side-effect findings suggest that among participants who said that the chairman did *not* intend to harm (63% of the group), a substantial portion would judge [1] to be correct. However, [1] was endorsed by only 26% of this group, making it again the least endorsed of the five statements (see Table 2).

People's more selective judgments of the most accurate and second-most accurate descriptions were striking. Only 2% of participants in the harm condition judged statement [1] as either the most accurate or second-most accurate description, well below chance levels (40% chance to select a given statement as most or second-most accurate), $\chi^2 = 81.4$, $p < .001$. In contrast, people selected statements [2] and [3] as most or second-most accurate at levels well above chance (88% and 68%, respectively), χ^2s > 45.0, ps < .001.

Although people tended not to view the "intentionally harmed" description as correct, they nonetheless assigned a substantial amount of blame to the agent (M = 4.77, SD = 1.34). Blame judgments correlated weakly with the dichotomous and multiple-response intentionality questions, both rs = .20.

Help condition. Among the multiple descriptions of the helping event, statement [1] was judged correct by 45% of participants, [5] by 48%, and [3] by 77%; none of these rates differed from the corresponding rates in the harm condition, all χ^2s < 1.0. However, there was one difference between the two conditions. Whereas in the harm condition endorsement of [1] was substantially lower (46%) than was endorsement of the dichotomous intentionality question (71%), Wilcoxon $z = 5.97$, $p < .001$, in the help condition the two endorsement rates were indistinguishable (45% and 51%, respectively), $p > .10$.

Among participants who said the chairman did not intend to help (55% of the group), only 29% endorsed [1]. Neither of these rates differed from the corresponding rates in the harm condition, both χ^2s < 1.5.

Statement [1] was judged as most or second-most accurate by 23% of people, below chance levels, $\chi^2 = 7.87$, $p < .01$. Statement [3] was most popular, selected by 89% of people, and this rate did not differ from the proportion selecting [3] in the harm condition (82%), $p > .10$.

Praise ratings were moderate (M = 3.33, SD = 1.27) but were substantially lower than blame ratings in the harm condition, $t(234) = 8.15$, $p < .001$, $d = 1.32$. Praise correlated weakly with the dichotomous ($r = .20$) and multiple-response intentionality questions ($r = .15$).

Discussion

Study 5 showed that when people are able to select any descriptions that correctly depict the chairman's behavior, few describe the act of harming (or helping) itself as intentional. The proportion of people in the harm condition who saw option [1] ("intentionally harmed") as correct (46%) was substantially smaller than the proportion of people who assented to the dichotomous intentionality question either in this study (71%), in Study 1 (87%), or in Knobe's (2003a) original study (82%), all χ^2s > 16.0, ps < .001. Furthermore, when we examined only the people who said that the harming was unintended, 26% endorsed the "intentionally harmed" option. Finally, when people selected the most or second-most accurate description, only 2% endorsed the "intentionally harmed" option. This is the most telling measure if we wonder what a jury's judgment might be in legal proceedings, because it must converge on a single description of the case in question.

The results in the modified help condition mirrored those in the harm condition, with one exception: people's consistency between dichotomous intentionality judgments and multiple-response judgments. In the help condition, the "any correct" endorsement of the response option "intentionally helped" was nearly identical (45%) to that of the dichotomous question (51%); when people answered "yes" to the dichotomous question, they really meant it. In the harm condition, the any correct endorsement of the option "intentionally harmed" (46%) was lower than the dichotomous question (71%). Around one third of the people who said "yes" to the dichotomous question did not seem to really mean it.

So how do people interpret negative side effects? Study 5 showed that most people refrained from seeing the chairman as intentionally harming the environment; instead, they overwhelmingly saw him as pursuing his primary goal (i.e., to adopt a profit-raising program) *while fully knowing* that harm would occur. Thus, people appear to distinguish between *intentionally* bringing about a side effect and *knowingly* doing so. Linguistic modifiers such as *knowingly*, *willingly*, or *intentionally* are often treated as (near) synonyms in the law (Levinson, 2005; Malle & Nelson, 2003). However, people may make fine distinctions between them because they highlight the presence of different components of intentionality (i.e., *knowingly* for belief, *willingly* for desire, and *intentionally* for intention). Thus, according to the extant interpretation of the side-effect findings, people see the chairman as *intentionally* bringing about harm, whereas we propose that people see the chairman as *knowingly* bringing about harm. Study 6 directly contrasted these two interpretations.

Study 6 also addressed a possible criticism of Study 5. People might prefer the statement "The chairman intentionally adopted a program he knew would harm [help] the environment" over the statement "The chairman intentionally harmed [helped] the environment" because the former is longer and more informative than the latter, even though both may be equally accurate. There are two reasons to doubt this account. First, since the latter (broader) statement entails the former (more specific) statement, people may actually prefer the latter if they viewed both as accurate but had to select one (as they had to in response to the "most accurate" question), but participants in Study 5 did not show this pattern. Second, and more important, even when people were invited to select as many statements as they believed to be correct, they still endorsed the "intentionally harmed" statement far less often than the "knew would harm" statement. Study 6 added an empirical test of this alternative account. All response options were of identical length so that any preference for the *knowingly* description over the *intentionally* description must be due to people's conceptual interpretation, not to superficial features of the descriptions.

Table 3. Percentage of People Selecting Each Description as Most Accurate (Study 6)

Description	Most accurate
"The CEO willingly harmed the environment."	12
"The CEO knowingly harmed the environment."	86
"The CEO intentionally harmed the environment."	1
"The CEO purposefully harmed the environment."	1

Study 6

Method

Participants were 101 undergraduate students who completed a one-page questionnaire as part of a larger computer-presented survey and received partial course credit in return. All participants read the standard harm vignette (except that "CEO" replaced the label "chairman of the board"), then answered a yes–no intentionality question, followed by a 0 to 5 blame rating and the request to select the "most accurate description of what the CEO did." Because of the legal significance of the comparison between *intentionally* and *knowingly*, we asked for a single selection, just as a jury member would be asked to provide. The four descriptions, along with the percentage of people selecting each as the most accurate, are shown in Table 3.

Results

Intentionality judgments on the yes–no question were in line with previous findings (73% said "yes"). Also as usual, people strongly blamed the CEO ($M = 4.3$), but blame correlated weakly with intentionality, $r = .25$. Most important, hardly anyone thought it was most accurate to say that the CEO "intentionally" or "purposefully" harmed the environment (1% each; see Table 3). Instead, 86% found it most accurate to say that the CEO "knowingly harmed the environment," and 12% found that he "willingly" did so.

Replication

We replicated these results with a different story content (an Air Force captain decides to bomb a weapons factory even though he is told there will be a number of civilian deaths), and we added "deliberately" to the list of possible descriptions. Thus, three out of the five options were variants of "intentionally," two of which ("purposely," "deliberately") would be appealing if one wanted not only to be consistent with one's answer to the initial dichotomous intentionality question (which 57% endorsed) but also to say something slightly different. Nonetheless, out of 100 participants, 84% indicated that the captain "knowingly" killed the civilians, and only 5% selected any of the three variants of "intentionally."

Discussion

When people are allowed to choose their own conceptualization of the debated side-effect story, they predominantly consider "knowingly harming" the most accurate description of the protagonist's behavior. Only 2 of 101 respondents regarded "intentionally harming" or its variant, "purposefully," as most accurate. This finding is particularly noteworthy because a strong majority initially endorsed intentionality in the dichotomous forced-choice response, but once given other options, they picked a different label as the most accurate description. Not even priming or pressures of consistency could persuade people to describe the protagonist as having intentionally harmed the environment.

One might wonder, though, whether people felt pressure to say something *different* to the multiple-choice question than they did to the dichotomous question. However, they did not merely pick *any* different option; almost all of them picked the "knowingly" option. Moreover, in a follow-up study, 30 participants read the standard CEO harm scenario and selected "most accurate" descriptions without first answering a dichotomous intentionality question. In that case, too, only 10% chose the "intentionally harmed" option as either the most or second-most accurate, and 83% chose the option corresponding to "knowingly."

How is it possible that almost nobody considered the "intentionally harmed" description to be accurate when, among these same responders, 73% acquiesced to the dichotomous question of whether the CEO intentionally harmed the environment?

Consider the three mental components of intentionality: belief, desire, and intention. The CEO knows (i.e., has a belief) that his action will bring about harm and, by dismissing the obligation to prevent harm, reveals a degree of desire for this harm (as shown in Studies 1-4). But there is little evidence for the third mental component: The CEO did not form an *intention* to harm the environment. This is why—when given a choice—most people judge that the CEO knowingly (with belief) or willingly (with desire) harmed the environment, but not that he intentionally harmed it.

The CEO does, however, intend to proceed with his primary goal (to make a profit), and he intentionally defies the norm to prevent harm because he adopts a program with a known harmful side effect. This intentional defiance likely makes it difficult for people to say "no" to the standard dichotomous intentionality question, because it would amount to declaring that the CEO "unintentionally" or "accidentally" harmed the environment. Once they are freed from the forced dichotomous choice and have a chance to select a more fine-grained interpretation of the situation, they almost uniformly pass over the "intentionally harmed" label in favor of characterizing the CEO's behavior as *knowingly* or *willingly* harming the environment.

Together, the results of Studies 5 and 6 demonstrate two points. First, people conceptually differentiate between the subtly different mens rea concepts of knowingly, willingly, and intentionally performing an action. In the absence of an agent's intention to harm, they characterize the agent as acting knowingly or willingly, but not intentionally. Second, people distinguish between primary intentional actions (e.g., adopting an economic program) and the side effects of those actions (e.g., harm to the environment). Despite many researchers' claims (Knobe, 2003a, 2004b; Nadelhoffer, 2006a, 2006b; Nichols & Ulatowski, 2007; Wright & Bengson, 2009), people's dominant interpretation does not view side effects themselves as intentional (particularly when they are unintended). Rather, side effects are seen as known consequences of an agent's primary action, and for knowingly bringing about such consequences, agents are blameworthy.

General Discussion

We have identified two factors that account for the surprising side-effect findings, in which people appeared to judge unintended negative, but not positive, side effects as intentional (Cokely & Feltz, 2009; Knobe, 2003a; Leslie et al., 2006; Nadelhoffer, 2006a). First, variations in perceived desire directly affect intentionality judgments, and the original scenarios of negative and positive side effects were not equated for the agent's desire (Study 1). This is because people interpret not caring about negative outcomes as evidence of moderate desire, but not caring about positive outcomes as evidence of virtually no desire (Study 2). This pattern holds even in the absence of moral considerations (Study 3). Weakening the evidence of desire for the negative side effect reduces intentionality judgments substantially—to 40% compared with the original 82% in Knobe's (2003a) scenario (Study 4a) and to below 30% in related scenarios (Mele & Cushman, 2007; Phelan & Sarkissian, 2008; also see footnote 4). Moreover, strengthening the evidence of desire for the positive side effect increases intentionality judgments to over 50% (Study 4b), effectively converging with the negative side effect in this range.

Second, the standard dichotomous intentionality question has obscured how people actually think about the agent's relation to the side effect. When people have multiple options to describe the agent's behavior, they differentiate clearly between the primary act (e.g., adopting a program) that was truly intentional and the side effect that was knowingly (for some, even willingly) brought about (Studies 5 and 6). To be sure, the agent deserves much blame for knowing about and disregarding the harm. But that is not the same as intentionally bringing it about, and people recognize the difference between the two.

Theoretical Implications

Our findings contradict two conclusions previously drawn from the side-effect findings. The first was that people judge unintended immoral side effects as intentional (Knobe, 2003a, 2004b; Nadelhoffer, 2006a, 2006c). Such judgments would challenge most models of intentionality (Adams, 1986; Malle & Knobe, 1997; Mele, 1992), according to which any intentional action must have been intended. Our studies suggest that these models have largely weathered the challenge. When given sufficient response options, few people characterized an unintended side effect (e.g., harm to the environment) as *intentional*; most indicated that the agent *knowingly* brought it about. Even in Study 5, when participants could endorse any descriptions they believed to be correct, just 26% of those who judged the negative side effect unintended considered it to be intentional, which was indistinguishable from the corresponding proportion in the help case (29%). Across both conditions, only 16% of the entire sample exhibited the answer pattern of "intentional yet unintended." Thus, the vast majority of people interpret (negative or positive) side effects either as (a) not intentional or (b) intentional but also intended.

The second conclusion that our results call into question is that "people's intuitions as to whether or not a behavior was performed intentionally can be influenced by their beliefs about the moral status of the behavior itself" (Knobe, 2004a, p. 270). We have shown that the original evidence for this conclusion confounded moral valence with desire. Since people are expected to prevent negative and foster positive outcomes, "not caring" about an outcome indicates greater desire when the outcome is negative than when it is positive (and this pattern holds for moral as well as nonmoral norm violations; see Study 3). Importantly, when we manipulated the agent's desire for harm or help, the side-effect asymmetry disappeared (Studies 4a and 4b). These findings show that considerations of an agent's desire for a side effect are far more important for the question of intentionality than are considerations of moral valence.

Thresholds in Judging Intentionality

Our results show that once we correct for differences in desire between the negative and positive side-effect cases, no intentionality difference remains. In Study 5, when endorsing any correct descriptions of the CEO scenario, 46% of people in the harm condition and 45% of people in the help condition said that the side effect was intentional. Among those who said the side effect was not intended, the proportions were again nearly identical—26% in the harm condition versus 29% in the help condition. However, we found that more desire was "needed" in the positive case—when intentionality rates converged around 50% in the harm and help conditions (Studies 4a and 4b), desire ratings were higher in the help case ($M = 3.67$) than in the harm case ($M = 2.14$).

Thus, although there is little possibility that valence influences intentionality judgments once we equate inferred desire, the required evidence for a desire inference might vary by valence—and so might the required evidence for other components of intentionality, such as belief or skill. Signal detection theory (SDT: Swets, Tanner, & Birdsall, 1961) provides a useful framework here, as it holds that perceivers use a subjective criterion, or threshold, when deciding whether stimuli belong to one category or another. In this case, the stimuli are behavioral evidence and the categories are the components of intentionality (e.g., desire, belief, and intention). People may have a more lenient threshold of accepting evidence for these components when judging negative behaviors than when judging positive behaviors (Jones & Davis, 1965), because the costs associated with making errors in judging each type of behavior differ (cf. Haselton & Nettle, 2006).

Within SDT's framework, perceivers can make two kinds of errors in judgments of intentionality—they can mistakenly judge an unintentional behavior to be intentional (a "false alarm") or an intentional behavior to be unintentional (a "miss"). For negative behaviors, misses are more costly than false alarms, since undetected negative intentional actions both escape punishment and are likely to promote future harmful behavior. For positive behaviors, false alarms are more costly, since they would heap praise upon undeserving people. Accordingly, people may have a more lenient threshold for judging a negative action intentional than judging a positive action intentional. Because such judgments are grounded in the critical components of intentionality, threshold differences would apply to the decision of whether the agent had, say, a certain belief or a certain desire. In the CEO case, for example, an expression of indifference about the environmental harm sufficed for many people as evidence for a desire, whereas even an expression of excitement about the environmental benefit did not suffice for many people as evidence for a desire.

The Norm of Prevention and a Blame–Praise Asymmetry

Intentionality judgments aside, one asymmetry that emerged reliably and strongly was people's praise and blame judgments for causing side effects. Specifically, people were more inclined to blame agents who caused negative side effects than to praise agents who caused positive side effects (Studies 1, 4a, 4b, and 5). Every social community benefits from maximizing positive social events and minimizing negative social events, and it therefore rewards people's efforts to bring about positive events and prevent negative events (Hamilton, Blumenfeld, Akoh, & Miura, 1990). If a positive outcome occurs without the agent actively trying to achieve it, the community will not reward the person's inaction, so it withholds praise. By contrast, if a negative outcome occurs without the person trying to *prevent* it, the community must discourage such inaction; because the person violated a *norm of prevention*, the community will dole out blame. Applied to the side-effect cases, the chairman in the original help scenario did not show any effort to bring about the positive outcome (it happened on its own), so he deserves little praise; the chairman in the harming scenario did not show any effort to prevent the negative outcome, so he deserves substantial blame (cf. Wright & Bengson, 2009).

The impact of asymmetric norms on allowing negative versus positive outcomes can be illustrated by the use of the action modifier *knowingly*, which proved so important in accurately characterizing the protagonist's action in the negative side-effect scenarios. This modifier is readily applied to negative actions (e.g., "He knowingly sold a stolen car" and "She knowingly infected others with the virus"), but it is not typically applied to positive actions (e.g., "He knowingly pulled the victim out of the water"). A brief archival exploration illustrates this point: Among the first 20 results of a Google search for "knowingly" (excluding dictionary definitions), we found that all of them modified negative actions (e.g., lying, infecting, taking steroids). Because of the norm of preventing harm, knowingly doing or allowing something that causes harm, even if one does not cause it intentionally, deserves blame and is linguistically marked to invite such blame. Knowingly doing or allowing something that happens to cause benefit, if one does not explicitly try to foster it, deserves no praise and is not further linguistically marked.

Accounting for Additional Findings

Our conclusions help explain a number of other recent findings in the literature on intentionality and blame. Mallon (2008) reported a negative-positive asymmetry in intentionality judgments by using vignettes in which the agent said: "I admit it would be good to harm the Australians [to help the orphanage] . . . but I don't really care about that." Whereas "not caring" about a negative outcome conveys a (mild) desire for the outcome, not caring about a positive outcome conveys a clear lack of desire. Thus, Mallon's asymmetry—like Knobe's (2003a)—was most likely due to differences in desire, not differences in morality.

Other researchers propose nonmoral explanations of the side-effect findings. Machery (2008) emphasizes differences in trade-offs: The harm case describes a trade-off because the agent incurs a cost (harm to the environment) to achieve a benefit (profit), but there is no trade-off in the help case, as both outcomes are positive (help to the environment, profit). Since people "think of costs as being intentionally incurred in order to reap some foreseen benefits" (Machery, 2008, p. 177), people see the act of harming—but not that of helping—as intentional. Uttich and Lombrozo (2010) emphasize norms, arguing that acts of norm violation are judged more intentional than acts of norm conformity. We agree with the claims of Machery and Uttich and Lombrozo, but our account helps explain why their findings obtain. As we have argued, indifference about costs or about norm violations indicates a desire for or approval of the outcome, whereas indifference about benefits or norm conformity indicates a clear lack of desire.

Consistent with our account, Phelan and Sarkissian (2008) showed that an agent may know about a negative side effect but nonetheless bring it about unintentionally. The critical condition for their finding, we propose (and the authors now argue, too; see Phelan & Sarkissian, 2009), is that the agent lacked desire for the negative outcome ("I feel terrible about increasing joblessness . . .").

Open Questions

There are many questions we have not addressed in the present studies. For example, the agent's primary motive or action may influence the assessment of side effects. In most studied scenarios, the agent pursues a socially undesirable or neutral goal (e.g., making profits). An explicitly positive goal (e.g., adopting a program to save jobs) would likely alter people's perceptions of the agent's desire for the negative side effect and, as we would predict, ascriptions of intentionality (at least in the forced-choice response format). Another question is whether the certainty of the negative side effect (i.e., "the program . . . *will* harm the environment") makes the agent's dismissal of any prevention attempts more objectionable. What if there is only a "chance that the environment would be harmed"? In this case, the agent's dismissal of the harmful side effects might be interpreted as a belief that the harm is unlikely to ensue, not as desire for the harm, so intentionality rates should decrease.

The studies we report, like all extant work on the side-effect findings, were exclusively vignette studies. Using vignettes allowed us both to compare our studies directly with previous findings and to reasonably approximate the courtroom situation, in which jurors are presented with a set of facts and must answer constrained questions ("Did the defendant have intent?"). Nonetheless, it will be important for future research to assess the cognitive timing of blame, praise, and intentionality judgments to better determine their causal order. We are currently conducting reaction-time studies to assess which judgment outraces the other, and so far intentionality appears to be the winner (Guglielmo & Malle, 2009).

Conclusion

Can unintended negative side effects be intentional? Knobe's (2003a) puzzling findings caused an avalanche of research into the disconcerting possibility that people's judgments of morality shape their judgments of intentionality—not, as is commonly assumed, the other way around. We have shown that although people strongly blame others for knowingly bringing about negative side effects, once they have a chance to flexibly express their interpretation of the events, people rarely see these unintended side effects as intentional. Our findings highlight people's capacity for complex, systematic, and relatively unbiased information processing about the intentionality of human behavior, and they cast serious doubt on the hypothesis that judgments of intentionality are guided by moral considerations.

Acknowledgments

The authors thank Mark Alicke, Holly Arrow, Scott Ewing, Sara Hodges, Joshua Knobe, Thomas Nadelhoffer, Steven Sloman, and Jonathan Weinberg for helpful comments on earlier versions of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

Notes

1. There are of course cases in which people act unintentionally yet receive blame (e.g., for their negligence), but the relevant intentionality judgment is still before an assessment of blame (see Guglielmo, Monroe, & Malle, 2009, for a more detailed model).
2. Knobe no longer views the influence of morality on intentionality as a biased process but rather sees it as a constitutive one—the morality of an action fundamentally guides people's inclination to see it as intentional (Knobe, 2010; Pettit & Knobe, 2009). Thus, rather than a model of blame → intentional, Knobe currently appears to endorse a model of negative → intentional. Our claims apply to either instantiation of Knobe's model, but we will generally refer to the former instantiation as it is most consistent with other claims in the literature.
3. The exact formulation of all vignette materials used in the current studies can be viewed online at http://research.clps.brown.edu/soccogsci/SE/
4. In another study, using very different content, we showed the same effect by manipulating the protagonist's dispositional attitudes. Participants read about Mayor Spires, who knew that attracting a franchise to her city would reduce funding for a program to feed the homeless. Spires had a dispositional attitude that would either be *favoring* this outcome (she had "little concern for the city's lower-class folks") or *opposing* it (she tended to care about "the problems and concerns of all her constituents, both upper- and lower-class"). Participants indicated whether she brought about the side effect (the homeless going hungry) *intentionally*, as well as whether she *intended* this outcome (which, according to the standard model of intentionality, should reflect variations in the desire component if belief is constant). Results showed that the side effect was more often judged intentional for the *favoring* mayor (51%) than the *opposing* mayor (26%), $\chi^2 = 12.21$, $p < .001$, $d = .53$. Moreover, this effect was mediated by judgments of whether the mayor intended to bring about the side effect.
5. In two additional conditions we expected to weaken evidence of the agent's desire for the side effect. In one, the chairman said, "I truly feel terrible about harming the environment . . ." In the other, the chairman thought to himself, "I feel terrible about harming the environment . . ." However, in both cases, desire ratings did not differ from those in the original harm condition, thus failing the manipulation check (this may be because the chairman's primary goal of making profit is still socially undesirable—we explore this issue in the General Discussion). Since people did not differentiate between any of the variants, we grouped them into a single harm condition.

References

Abbey, A. (1987). Perceptions of personal avoidability versus responsibility: How do they differ? *Basic and Applied Social Psychology, 8*, 3-19.

Adams, F. (1986). Intention and intentional action: The simple view. *Mind and Language, 1*, 281-301.

Adams, F., & Steadman, A. (2007). Folk concepts, surveys, and intentional action. In C. Lumer & S. Nannini (Eds.), *Intentionality, deliberation, and autonomy: The action-theoretic basis of practical philosophy* (pp. 17-23). Aldershot, UK: Ashgate.

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*, 368-378.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*, 556-574.

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

Cokely, E. T. & Feltz, A. (2009). Individual differences, judgment biases, and theory-of-mind: Deconstructing the intentional action side effect asymmetry. *Journal of Research in Personality, 43*, 18-24.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy, 60*, 685-700.

Davis, C. G., Lehman, D. R., Silver, R. C., Wortman, C. B., & Ellard, J. H. (1996). Self-blame following a traumatic event: The role of perceived avoidability. *Personality and Social Psychology Bulletin, 22*, 557-567.

Fincham, F., & Jaspars, J. (1979). Attribution of responsibility to the self and other in children and adults. *Journal of Personality and Social Psychology, 37*, 1589-1602.

Guglielmo, S., & Malle, B. F. (2009, June). *The timing of blame and intentionality: Testing the moral bias hypothesis.* Poster presented at the annual meeting of the Society for Philosophy and Psychology, Bloomington, IN.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 52,* 449-466.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814-834.

Hamilton, V. L., Blumenfeld, P. C., Akoh, H., & Miura, K. (1990). Credit and blame among American and Japanese children: Normative, cultural, and individual differences. *Journal of Personality and Social Psychology, 59*, 442-451.

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 10,* 47-66.

Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: Wiley.

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, NY: Academic Press.

Kashima, Y., McKintyre, A., & Clifford, P. (1998). The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology, 1*, 289-313.

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. (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. (2005). Cognitive processes shaped by the impulse to blame. *Brooklyn Law Review, 71*, 929-937.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33,* 315-329.

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.

Levinson, J. D. (2005). Mentally misguided: How state of mind inquiries ignore psychological reality and overlook cultural differences. *Howard Law Journal, 49,* 1-29.

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

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3,* 23-48.

Malle, B. F. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture, 6*, 61-86.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101-121.

Malle, B. F., & Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 45-67). Cambridge, MA: MIT Press.

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 versus Machery: Testing the trade-off hypothesis. *Mind and Language, 23*, 247-255.

Mele, A. R. (1992). *Springs of action: Understanding intentional behavior*. New York, NY: Oxford University Press.

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

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

Nadelhoffer, T. (2004). The Butler problem revisited. *Analysis, 64*, 277-284.

Nadelhoffer, T. (2005). Skill, luck, control, and intentional action. *Philosophical Psychology, 18*, 341-352.

Nadelhoffer, T. (2006a). Desire, foresight, intentions, and intentional actions: Probing folk intuitions. *Journal of Cognition and Culture, 6*, 133-157.

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

Nadelhoffer, T. (2006c). On trying to save the simple view. *Mind and Language, 21*, 565-586.

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

Ohtsubo, Y. (2007). Perceiver intentionality intensifies blameworthiness of negative behaviors: Blame-praise asymmetry in intensification effect. *Japanese Psychological Research, 49*, 100-110.

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

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

Pizarro, D. A., Laney, C., Morris, E. K., & Loftus, E. F. (2006). Ripple effects in memory: Judgments of moral blame can distort memory for events. *Memory and Cognition, 34*, 550-555.

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology, 83,* 789-803.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind.* Cambridge, UK: Cambridge University Press.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York, NY: Springer.

Shultz, T. R., & Wells, D. (1985). Judging the intentionality of action-outcomes. *Developmental Psychology, 21*, 83-89.

Solan, L. M. (2003). Cognitive foundations of the impulse to blame. *Brooklyn Law Review, 68,* 1003-1028.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68*, 301-341.

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

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York, NY: Guilford.

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