Understanding Side-Effect Intentionality Asymmetries: Meaning, Morality, or Attitudes and Defaults?

Personality and Social
Psychology Bulletin
1-16

DOI: 10.1177/0146167220928237
journals.sagepub.com/home/psp
SAGE

Sean M. Laurent[1], Brandon J. Reich[2], and Jeanine L. M. Skorinko[3]

Abstract

People frequently label harmful (but not helpful) side effects as intentional. One proposed explanation for this asymmetry is that moral considerations fundamentally affect how people think about and apply the concept of intentional action. We propose something else: People interpret the meaning of questions about intentionally harming versus helping in fundamentally different ways. Four experiments substantially support this hypothesis. When presented with helpful (but not harmful) side effects, people interpret questions concerning intentional helping as literally asking whether helping is the agents' intentional action or believe questions are asking about why agents acted. Presented with harmful (but not helpful) side effects, people interpret the question as asking whether agents intentionally acted, knowing this would lead to harm. Differences in participants' definitions consistently helped to explain intentionality responses. These findings cast doubt on whether side-effect intentionality asymmetries are informative regarding people's core understanding and application of the concept of intentional action.

Keywords

intentionality, intentionality asymmetries, side-effect effect, Knobe effect, moral judgments

Received May 23, 2019; revision accepted April 29, 2020

From infants to elders, people perform many intentional actions each day, hoping to achieve outcomes that satisfy desired goals. Although most actions are not noteworthy, understanding others' behavior is sometimes vitally important. This helps explain people's keen interest in explaining intentional actions (Malle & Knobe, 1997b) and why the ability to decipher intentionality of behavior is the heart of social cognition (Malle et al., 2001).

Because people typically have reasons for acting, people may assume that most actions are intentionally performed (Rosset, 2008): Workers work to earn paychecks; students study to pass classes; people drink to quench thirsts. Yet, although intentionality inferences are rapid (Malle & Holbrook, 2012) and seemingly "easy," they necessitate, are accompanied by, or result in multiple inferences about agents' mental states. For example, they imply beliefs that agents desired particular outcomes, intended to act, were aware of acting, and thought their actions would result in these desired outcomes (Malle & Knobe, 1997a).

Intentional actions often have further effects beyond helping agents achieve intended goals. Usually, these side-effects-of-actions are inconsequential, unknown to and unanticipated by actors. Other times, they demand attention. For example, when goal-directed actions lead to harm, perceivers will try to infer an acting agent's mental states: Was the harm foreknown or should it have been? Did the agent desire the outcome? If so, perceivers might respectively label behaviors as reckless or negligent (e.g., Alicke, 2008; Fitzgerald & Williams, 1962; Laurent, Nuñez, & Schweitzer, 2015; Laurent et al., 2016; Nuñez et al., 2014; Stark, 2016) and deserving of censure.

The Side-Effect Effect (SEE)

Importantly, when goal-directed actions lead to foreknown harmful side effects to which agents are indifferent, people frequently label them as intentional, even though they are not particularly desired or intended. Yet, when the same goal-directed actions lead to helpful side effects, people rarely

[1] University of Illinois Urbana–Champaign, USA
[2] Portland State University, OR, USA
[3] Worcester Polytechnic Institute, MA, USA

Corresponding Author:
Sean M. Laurent, Department of Psychology, University of Illinois Urbana–Champaign, Champaign, IL 61820, USA.
Emails: seanmlaurent@gmail.com; slauren@illinois.edu

label them as intentional. Consider the following, widely reported in the literature:

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but it will also [help] [harm] the environment." The chairman of the board answered, "I don't care at all about [helping] [harming] the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was [helped] [harmed]. (Knobe, 2003)

Most people respond "yes" when asked, "Did the chairman intentionally *harm* the environment?" Yet, most people respond that he did not intentionally *help*. This asymmetry in whether people label side effects as intentional has been called the SEE (e.g., Cova & Naar, 2012), and has been difficult to fully explain (e.g., Nichols & Ulatowski, 2007). One proposed explanation is that moral considerations pervasively influence people's intuitions about what constitutes intentional action (e.g., Pettit & Knobe, 2009). However, this reverses conventionally held beliefs about the causal ordering of intentionality inferences and moral judgments (e.g., Malle et al., 2014), challenging the idea that intentionality inferences require agents to explicitly desire and intend outcomes (e.g., Malle & Knobe, 1997a; cf. Knobe, 2003). This suggests that understanding the SEE is theoretically important. It is also practically important because intentionally caused harms are punished more severely than negligent and reckless harms (e.g., Darley & Pittman, 2003). Thus, agents who cause side-effect harm might be punished like those who directly intend harm. Another proposed explanation is that multiple definitions of intentionality exist, and that outcome valence (e.g., beneficial vs. harmful side effects) can influence which concept is applied (e.g., Cova et al., 2012; Cushman & Mele, 2008; Lanteri, 2012; Mele & Cushman, 2007). Thus, determinations regarding whether help or harm was intentional might rely on different meanings of intentionality in different cases.

Although either proposition is possible, explaining the effect may not require multiple definitions of intentionality or that moral considerations influence intuitions about intentionality. The current article investigates an alternative hypothesis: that people understand questions about the intentionality of side effects to be asking different things in the two cases and that responses across conditions are therefore to substantively *different questions*. If true, this would make it difficult to draw conclusions regarding ordinary understanding of intentional action based on comparisons of responses across the two cases. To test our hypothesis and compare it with other propositions, this work empirically explores what best explains intentionality responses: definitional differences (as we propose), moral judgments, or the position of agents' perceived attitudes toward helping/harming relative to different moral defaults.

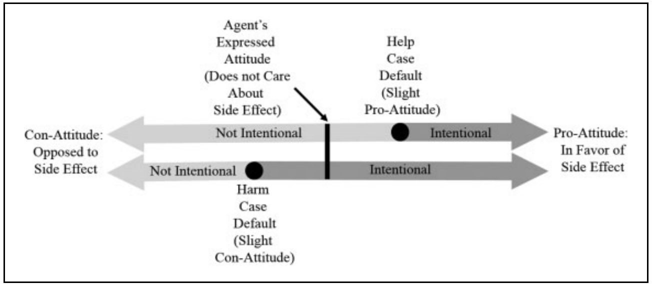
Agent's Expressed Attitude (Does not Care About Side Effect)
Help Case Default (Slight Pro-Attitude)
Con-Attitude: Opposed to Side Effect
Not Intentional
Intentional
Pro-Attitude: In Favor of Side Effect
Not Intentional
Intentional
Harm Case Default (Slight Con-Attitude)

Figure 1. Graphical representation of Knobe's (2010; Pettit & Knobe, 2009) explanation of intentionality side-effect asymmetries.

The Moral Influence Hypotheses

Interest in the SEE has been broad because it suggests "that people's intuitions about whether an outcome was *intentionally* [italics in original] produced seem to vary depending on the moral status of the outcome itself" (Nichols & Ulatowski, 2007, p. 346), or that "moral evaluations might shape our understanding of actions as intentional or not" (Cova et al., 2016, p. 1295). Yet, intentions and intentionality have traditionally been assumed—whether correctly or not—to cause rather than follow social and moral evaluation (e.g., Hamlin et al., 2007; Vaish et al., 2010; see also Malle et al., 2014).

One explanation for the asymmetry in intentionality responses is that the moral valence of outcomes (i.e., how "bad" outcomes seem) directly influences intuitions about intentional action. Thus, one might hypothesize that to the extent perceivers believe agents' actions are wrong/forbidden, the more likely they will label the bad side effects these actions produce as intentional. A more elaborated proposal (Knobe, 2010; Pettit & Knobe, 2009) suggests the effect arises from the position of agents' attitudes toward side effects relative to normative moral defaults, or what one might reasonably expect the agents' attitude to be (Figure 1). For good side effects, one might expect most people to be at least slightly pro-helping. For bad side effects, most people should be at least slightly con-harming. Defaults are proposed to be somewhat decontextualized, requiring no evaluation of acting agents or their observed behaviors. However, determinations regarding the intentionality of side effects do require evaluating agents and their behaviors. In both cases, agents directly state that they do not care about the side effect. Thus, in help cases, agents' attitudes fail to be pro-helping enough (below the default) to be viewed as intentional. In harm cases, they are not con-harming enough (above the default; see Figure 1).

This hypothesis deserves further consideration. First, agents' attitudes are treated as a constant. If so, what matters most should be perceived defaults. However, despite being told otherwise, participants may infer attitudes other than indifference. Another issue regards the defaults. Depending on the case, defaults representing perceivers' *personal* opinions about what agents' attitudes should be (i.e., prescriptive judgments) might typically be that the agent should be very strongly or completely opposed to harming and fully in support of helping, particularly when helping is cost-free (Machery, 2008). In this case, it is unlikely that people would ever view agents' attitudes as "extreme enough" to surpass defaults and make harming seem unintentional or helping intentional. Thus, a better approach might be to assess participants' understanding of relevant descriptive norms, or what they think most people would do in the agent's situation. In any case, it is unclear which elements should be most important to intentionality asymmetries: perceived attitudes, defaults, or attitudes relative to defaults.

The Question Interpretation Hypothesis

Another proposition, beyond one suggesting multiple definitions of intentionality, is that the intentionality asymmetry arises from people interpreting the meaning of questions about the intentionality of side effects differently when they are harmful versus helpful (Adams & Steadman, 2004; Laurent et al., 2019). For example, because prescriptive and proscriptive morality differ (e.g., Janoff-Bulman et al., 2009), people evaluate morally praiseworthy and blameworthy behaviors in different ways (e.g., Guglielmo & Malle, 2019). Recklessness—typically defined as causing unintended but foreknown harm (Garner, 2014)—is also psychologically "close" to intentionality (Darley & Pittman, 2003) and has no analogue for behaviors leading to positive outcomes. Moreover, like "recklessly," the word "intentionally" is typically invoked to reference negative, blameworthy outcomes (Malle, 2006).

This may explain why, in side-effect cases, people prefer "knowingly harmed" over other options when this choice is offered (e.g., Guglielmo & Malle, 2010). However, participants are typically asked about intentional harming rather than foreknown harming, suggesting that questions may not be taken at face value. Instead, people may reinterpret questions about side-effect harm to be asking whether the agent's goal-directed behavior (e.g., "starting a program to increase profits") led to foreknown harm. In help conditions, the same question, with only one word changed, may prompt a more straightforward interpretation (e.g., "was helping what the agent intentionally did?") or be perceived as asking whether an agent's goal was to help. Likewise, probably because harming agents seem quite blameworthy, but helping agents do not deserve praise, people may believe that responding "no" in harm cases would sound exculpatory and responding "yes" in help cases would suggest deservingness of praise (Laurent et al., 2019).

We note that Cova and colleagues (2016) found that rates of labeling harming as intentional did not significantly differ when comparing cases where participants responded to standard intentionality questions versus when additional options (e.g., willingly or knowingly [harmed]) were also allowed. This might suggest that participants in harm conditions are not reinterpreting the intentionality question but applying multiple definitions. However, it is difficult to draw firm conclusions from null effects. Still, past work shows that when the choice is given, many participants indicate that an agent has both intentionally *and* knowingly harmed (Adams & Steadman, 2007). Potentially reconciling this, participants may interpret questions about intentional harming to mean agents "intentionally *did something, knowing* it would lead to harm." Findings from Cova and colleagues (2016) cannot rule this possibility out because (a) this option was not offered, (b) participants were not asked how they interpreted the intentionality question, and (c) participants could select multiple statements as correct, rather than choosing one statement that *best* described their opinion.

Overview of Experiments

Four experiments are reported that directly test the primary prediction that people exposed to helping and harming side-effect stories will define questions about intentionally helping/harming in different ways. A secondary hypothesis is that definitional differences will statistically mediate intentionality asymmetries more strongly than other potential mechanisms. In Experiments 1 and 2, we also examine whether moral judgments regarding the wrongness of behavior or of the agents' deservingness of praise/blame help explain the pattern of intentionality responses. In Experiments 3 and 4, we also investigate perceptions of the agents' attitudes toward the side effects, perceptions of default attitudes, and comparisons of perceived attitudes to defaults, exploring whether the latter mediates intentionality responses.

Participants in all experiments were recruited through Amazon's MTurk website, paid a small fee for their participation, and after providing consent, participated in only one condition of one experiment. Sample sizes for all experiments were determined in advance and no analyses were performed until target sample sizes were reached. With $\alpha = .05$ (two-tailed), sample sizes had 80% power to detect effect sizes from $d = 0.32$ (Experiment 4) to $d = 0.40$ (Experiment 2). All manipulations and measures are disclosed, and all procedures, vignettes, and exact wording of measures can be found in the Supplemental Appendix. Analyses including (vs. excluding) the few participants who failed to pass one or more attention check question did not substantively differ, so all participants were retained. De-identified data for all experiments are freely available in the Supplemental Appendix of this article.

Experiment 1

Experiment 1 used the original chairman vignette (Knobe, 2003), described earlier. Participants first indicated whether the chairman intentionally helped/harmed the environment.

Table 1. Means, *SD*, *t*, *p*, *d*, and 95% CI for Experiment 1.

Dependent Variable	Help *M* (*SD*)	Harm *M* (*SD*)	*t*(199)	*p*	*d*	95% CI $M_{\text{Difference}}$
Literal vs. knowledge	−1.81 (2.77)	2.57 (2.08)	12.60	<.001	1.79	[3.69, 5.06]
Goal vs. knowledge	−1.82 (2.70)	1.82 (2.76)	9.40	<.001	1.33	[2.87, 4.40]
Desire	–2.69 (2.02)	0.82 (1.90)	12.63	<.001	1.79	[2.96, 4.06]
Praise/blame	–2.59 (2.23)	3.67 (1.35)	23.91	<.001	3.40	[5.75, 6.78]
Immoral act	–2.80 (1.81)	3.11 (1.37)	25.93	<.001	3.68	[5.46, 6.26]

Note. t, d, and CI are reported as positive values. CI = confidence interval.

Next, they responded to items regarding whether they understood the intentionality question to be literally asking whether his intentional action was to help/harm versus asking whether he intentionally acted, knowing the action would lead to side-effect help/harm. In addition, participants reported whether the chairman's goals or foreknowledge were more important to their intentionality decisions, whether he wanted to help/harm the environment, whether he, respectively, deserved any praise/blame for the side effect in help/harm conditions, and whether his actions were wrong and should not have been performed.

We expected the typical asymmetry in intentionality responses to replicate. Our primary hypotheses were that harm condition participants would interpret the intentionality question as asking whether the agent intentionally acted, knowing harm would result and would indicate that their intentionality responses were more influenced by his foreknowledge than his goals. In the help condition, we expected participants to understand the question as asking whether the chairman's intentional action was to help and to focus more on his goals than his foreknowledge. We also expected that harm (vs. help) condition participants would perceive the chairman as having greater desire for the outcome, believe he deserved more blame than the helping chairman deserved praise, and see his actions as substantially more wrong. Finally, we hypothesized that the meaning question would mediate the effects of condition on intentionality responses. We remained agnostic as to whether other variables would mediate the effect.

Method

Participants. One hundred ninety-nine participants completed the study (46.7% female, 51.3% male, 2% other/did not disclose; $M_{\text{age}} = 32.33$, $SD = 11.27$). Participants leaned slightly liberal on a scale running from 1 = *extremely liberal* to 7 = *extremely conservative* ($M = 3.83$, $SD = 2.14$).

Procedure and measures. Participants first read the help or harm version of the original chairman vignette. This was followed by several questions. Intentionality: "Did the chairman intentionally help/harm (HH) the environment?" (0 = no, 1 = yes). Literal versus knowledge: "What best captures the meaning of the question about intentionality that you answered? The question is asking whether the..." (–4 = chairman's intentional action was literally to HH the environment, 4 = chairman intentionally did something, knowing the environment would be HH).[1] Goal versus knowledge: "What was more important to you in making your decision about whether the chairman intentionally HH the environment?" (–4 = the chairman's intention/goal, 4 = the chairman's knowledge that the environment would be HH).[2] Desire: To what extent did the chairman want to HH the environment?" (–4 = the chairman did not want to HH the environment, 4 = the chairman wanted to HH the environment). Praise/blame: "To what extent does the chairman deserve to be praised/blamed for HH the environment?" (–4.5 = the chairman deserves no praise/blame at all, 4.5 = the chairman deserves a lot of praise/blame). Immoral act averaged two items ($r = .91$): "Given that he knew the environment would be HH, what the chairman did was wrong," and "The chairman should not have done what he did if he knew it would also HH the environment." (–4 = completely disagree, 4 = completely agree).

Results

Intentionality responses. In the help condition, few participants labeled helping as intentional (16/100). In the harm condition, most participants labeled harming as intentional (93/99), $\chi^2(1) = 121.99$, $p < .001$, $\varphi = .78$.

Mean differences. Condition-based differences for remaining variables were first examined using independent-samples *t*-tests with 197 *df*. These results, summarized in Table 1, show that participants in the harm condition tended to interpret the intentionality question as asking whether the chairman intentionally acted (e.g., starting a program), knowing this would lead to the environment being harmed. In the help condition, participants interpreted the question more as asking whether his intentional action was literally to help. Harm condition participants also indicated that the chairman's foreknowledge was more important to their intentionality responses than his goal, that he desired the outcome somewhat, and that he deserved substantial blame for the outcome. In the help condition, participants believed the chairman's goal (vs. foreknowledge) was more important to their intentionality responses, that he did not desire the

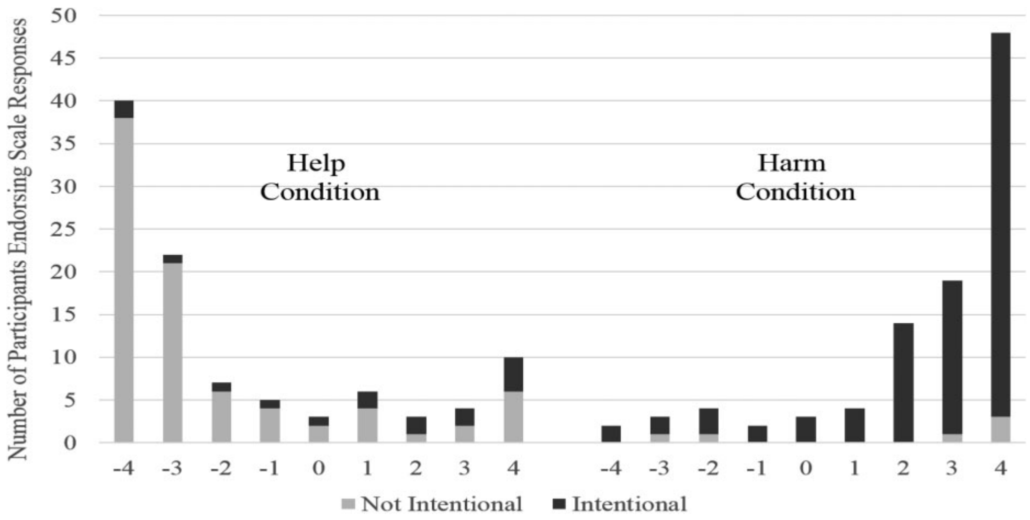

Figure 2. Frequencies of labeling helping or harming as intentional as a function of responses to the literal versus knowledge question in Experiment 1.

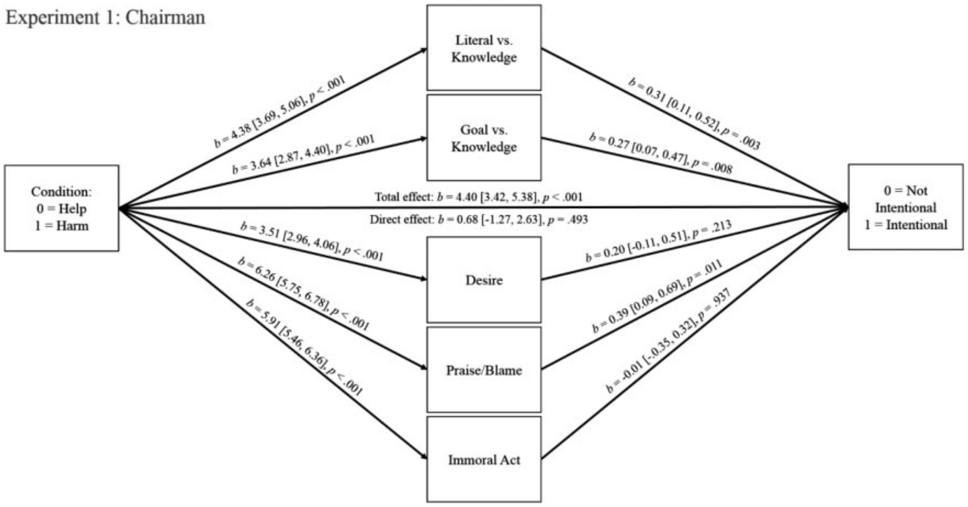
Experiment 1: Chairman
Condition: 0 = Help 1 = Harm
Literal vs. Knowledge
Goal vs. Knowledge
Desire
Praise/Blame
Immoral Act
0 = Not Intentional 1 = Intentional
b = 4.38 [3.69, 5.06], p < .001
b = 3.64 [2.87, 4.40], p < .001
Total effect: b = 4.40 [3.42, 5.38], p < .001
Direct effect: b = 0.68 [-1.27, 2.63], p = .493
b = 3.51 [2.96, 4.06], p < .001
b = 6.26 [5.75, 6.78], p < .001
b = 5.91 [5.46, 6.36], p < .001
b = 0.31 [0.11, 0.52], p = .003
b = 0.27 [0.07, 0.47], p = .008
b = 0.20 [-0.11, 0.51], p = .213
b = 0.39 [0.09, 0.69], p = .011
b = -0.01 [-0.35, 0.32], p = .937

Figure 3. Unstandardized path coefficients with 95% confidence intervals for mediation model in Experiment 1.

outcome, and deserved little praise. Importantly, harm condition participants viewed the chairman's action as quite immoral but help condition participants did not. Within both conditions, all means significantly deviated from scale midpoints (ps < .001).

Frequencies of literal versus knowledge responses. In Figure 2, we provide the frequencies of participants who labeled helping or harming as intentional as a function of endorsing particular responses to the literal-versus-knowledge definition question. Most help condition participants (62/100) strongly endorsed (i.e., either a -4 or -3 on the 9-point scale) that the intentionality question was literally asking whether the chairman's intentional action was to help. In contrast, 67/99 harm condition participants strongly endorsed that the intentionality question was asking whether the chairman intentionally acted, knowing harm would result.

Mediation. Indirect effects were tested using the process macro (Hayes, 2013), which allows dichotomous dependent variables and multiple simultaneous mediators. Standard errors of indirect effects were generated using bias-corrected bootstrapping (10,000 resamples) to create 95% confidence intervals (CIs). Condition was used to simultaneously predict all variables, with each mediator also used to predict intentionality responses (Figure 3). Indirect effects were significant through literal versus knowledge (ab = 1.38, 95% CI, [0.19, 2.76]) and goal versus knowledge (*ab* = 0.98, 95% CI = [0.17, 2.04]). Indirect effects through desire (*ab* = 0.69, 95% CI = [–1.16, 2.62]), praise/blame (*ab* = 2.43, 95% CI = [–0.22, 4.75]), and immoral act (*ab* = –0.08, 95% CI = [–2.61, 1.97]) were not significant, as 95% CIs contained zero. The direct effect of condition on intentionality was not significant.

Experiment 1 Replication

One possibility is that participants might not have responded to the "literal versus knowledge" question—which probed how participants understood the intentionality question—in the same way if the "literal" end of the scale had asked whether the chairman intentionally helped/harmed, rather than having asked whether his intentional *action* was to help/harm.[3] Experiment 1 was therefore replicated (N = 201),[4] using the same dichotomous intentionality question and a new literal versus knowledge question that asked, "What do you think best captures the meaning of the (intentionality) question? The question is . . . " (–4 = literally asking if the chairman intentionally helped/harmed the environment, 4 = asking whether the chairman intentionally did something [i.e., started a program], knowing that the environment would be helped/harmed). Results of this experiment were entirely consistent with those of Experiment 1 (i.e., significant between-condition differences on this question [p < .001, d = 0.81] and a significant indirect effect [p < .05] on intentionality responses through the meaning question), suggesting "intentional action was to help/harm" and "intentionally helped/harmed" may be understood in similar ways. Full results are available in the Supplemental Appendix of this article.

Discussion

Across help and harm conditions, participants indicated that the question "Did the chairman intentionally help/harm the environment?" was asking different things. This difference in understanding mediated intentionality responses, whereas other potential mechanisms did not. Most help condition participants interpreted the intentionality question as asking whether the chairman's intentional action was literally to help. In contrast, most harm condition participants indicated that the question was asking whether the chairman intentionally did something (i.e., started a program), *knowing* this would lead to harm. This suggests one explanation for the side-effect intentionality asymmetry. To the extent that help condition participants interpret a question about intentional helping as literally asking whether the agent's intentional action was to help, they may respond "no, not intentional" because helping was only a side effect of his intentional action. In harm conditions, to the extent that participants think the question is asking whether agents intentionally did something, knowing their action would lead to harm, participants may respond "yes, intentional" because the agents did intentionally act, foreseeing harm.

Most help condition participants also indicated that the chairman's goals (i.e., increasing profits) were more important than his foreknowledge for their responses to the intentionality question. In the harm condition, this was reversed: Foreknowledge was viewed as more important than goals, which accords with their described interpretation of the intentionality question. Responses to this question also significantly mediated intentionality responses, probably for a similar reason. We argue that this fairly straightforward vignette—which provides a wealth of information about an agent who acts for one reason, knowing but not caring that this will lead to incidental harm or benefit—promotes relatively complex processing of the available information (e.g., Royzman & Hagan, 2017), leading to the observed differences in understanding of the intentionality question and beliefs about what is important to consider when responding to it. Ultimately, differences in the relevance of the agent's goal versus foreknowledge across conditions may drive differences in intentionality responses.

Significant between-condition differences were also found for perceived desire to help/harm and judgments about whether the chairman deserved praise/blame. However, neither of these variables mediated intentionality responses. In addition, the variable associated with the largest effect size asked about the extent to which the chairman's actions were perceived as wrong or proscribed. Strong moral inferences were clearly drawn about his behavior but, at least in the help case, appeared to be distinct from judgments of praiseworthiness. That is, in the help condition, although people believed the agent did nothing wrong, they did not believe he deserved praise, probably because helping was not the goal motivating his action. In fact, given his stated attitude, it appeared that he did not *want* to help, even though helping was cost-free. This is clearly counter-normative. In the harm condition, though, the agent was viewed as doing something quite wrong and as deserving substantial blame because even though he did not particularly want to harm, he knew harm would result from his actions and acted anyway. Yet, despite the strong effect of condition on moral judgments, moral judgments failed to mediate intentionality responses, suggesting little direct role in causing the observed asymmetries in intentionality responses. That is, although the most "important" conclusions being drawn in side-effect cases may involve morality, moral judgments do not appear to directly shape whether people label actions as intentional or not. Moreover, to the extent that moral judgments play a role, it may be in shaping how people interpret questions about intentionality rather than by influencing intuitions about intentional action.

Experiment 2

Experiment 2 replicated Experiment 1 using a side-effect scenario adapted from Nadelhoffer (2006; see also Laurent,

Table 2. Means, *SD*, *t*, *p*, *d*, and 95% CI for Experiment 2.

Dependent Variable	Help *M* (*SD*)	Harm *M* (*SD*)	*t*(196)	*p*	*d*	95% CI $M_{\text{Difference}}$
Literal vs. knowledge	−1.89 (2.81)	1.61 (2.99)	8.48	<.001	1.21	[2.69, 4.31]
Goal vs. knowledge	–2.00 (2.70)	0.92 (3.19)	6.94	<.001	0.99	[2.09, 3.75]
Desire	−1.84 (2.16)	−1.27 (2.13)	1.87	.064	0.27	[–0.03, 1.17]
Praise/blame	–0.09 (2.31)	2.51 (1.83)	8.78	<.001	1.25	[2.01, 3.18]
Immoral act	–2.89 (1.80)	3.14 (1.38)	26.50	<.001	3.68	[5.58, 6.48]

Note. t, d, and CI are reported as positive values. CI = confidence interval.

Clark, & Schweitzer, 2015), where a bear-hunter fires a gun at a large bear to win a contest, knowing that this will help/ harm a nearby birdwatcher. Like the chairman, the hunter does not care about the side effect.

Method

Participants, procedure, and measures. One hundred ninety-eight people participated (88 female and 104 male participants, six undisclosed; $M_{\text{age}} = 32.76$, $SD = 11.55$, $M_{\text{ideology}} = 4.03$, $SD = 2.19$ on the same 7-point scale used in Experiment 1). Aside from using a different vignette, procedure and measures were identical to Experiment 1; however, praise/ blame was measured on a 9-point scale. The correlation between immoral act items was high ($r = .90$).

Results and Discussion

Intentionality responses. Few help condition participants labeled helping as intentional (16/97). A majority of harm condition participants labeled harming as intentional (69/101), $\chi^2(1) = 54.24$, $p < .001$, $\varphi = .53$.

Mean differences. Independent samples *t* tests (196 *df*) showed that with the exception of desire, significant differences emerged on all variables (see Table 2). With the exception of praise in the help condition ($t = 0.40$, $p = .694$), within both conditions, means significantly differed from scale midpoints, *p*s < .001 (for goal vs. knowledge in the harm condition, $p = .005$). On average, harm condition participants indicated that the intentionality question was asking whether the hunter intentionally shot at a bear to win a contest, knowing it would lead to harm. Help condition participants interpreted the question more as asking whether helping was literally the hunter's intentional action. Similarly, harm (vs. help) condition participants thought foreknowledge was more important than goals to their intentionality responses, thought the chairman deserved more blame (than praise), and thought the chairman's behavior was more immoral.

Frequencies of literal versus knowledge responses. A majority of participants in the help condition again strongly endorsed that the intentionality question was asking whether the hunter's intentional action was to literally help the birdwatcher. In the harm condition, a majority strongly believed the question was asking whether he intentionally acted, knowing the birdwatcher would be harmed (Figure 4).

Mediation. Indirect effects (see Figure 5) were significant through literal vs. knowledge ($ab = 1.69$, 95% CI = [0.88, 2.60]), goal versus knowledge ($ab = 1.26$, 95% CI = [0.52, 2.25]), and praise/blame ($ab = 1.55$, 95% CI = [0.37, 3.07]). Indirect effects through desire ($ab = 0.25$, 95% CI = [–0.03, 0.76]) and immoral act ($ab = 0.87$, 95% CI = [–1.48, 5.09]) were not significant. The direct effect of condition on intentionality was not significant, $c' = 0.02$, 95% CI = [–2.01, 2.04].

Discussion

Results of Experiment 2 replicated those of Experiment 1 using a different vignette. Participants again tended to define the intentionality question in different ways across conditions and relied on different mental states to answer the question. However, unlike in Experiment 1, praise/blame judgments mediated the effects of condition on intentionality responses. Although Knobe (2010) directly mentions that blameworthiness of agents for outcomes is not part of the described theoretical account, it may play some role (e.g., Alicke, 2008; Alicke & Rose, 2010). That is, although consideration of praise/blame may not be required to generate side-effect intentionality asymmetries (Knobe & Mendlow, 2004), in situations where it is relevant, participants may believe intentionality questions are in part asking about praise or blame, helping explain the effect (Laurent et al., 2019). However, as this variable did not mediate intentionality responses in Experiment 1, this finding should be interpreted cautiously.

Experiment 3

Experiment 3 again used the original chairman vignette. Rather than asking about moral judgments of the action as in Experiments 1 and 2, we asked several new questions regarding perceptions of the agent's pro-con attitudes toward the outcome (i.e., whether he was in favor of vs. opposed to it), what his attitudes *should have* been (i.e., personal

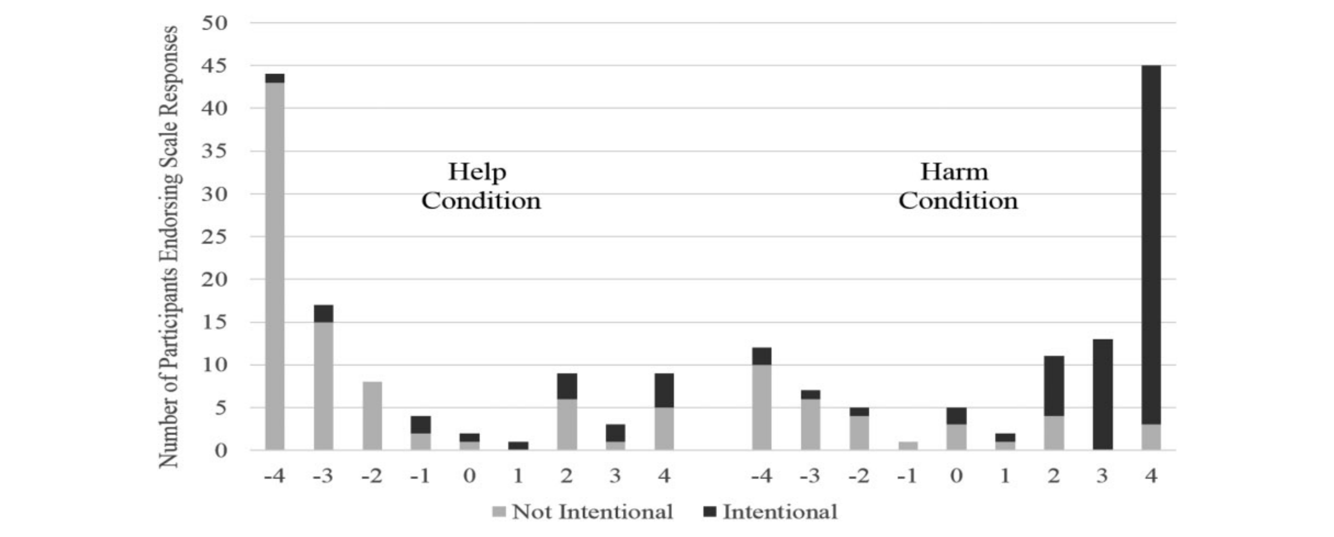

Figure 4. Frequencies of labeling helping or harming as intentional as a function of responses to the literal versus knowledge question in Experiment 2.

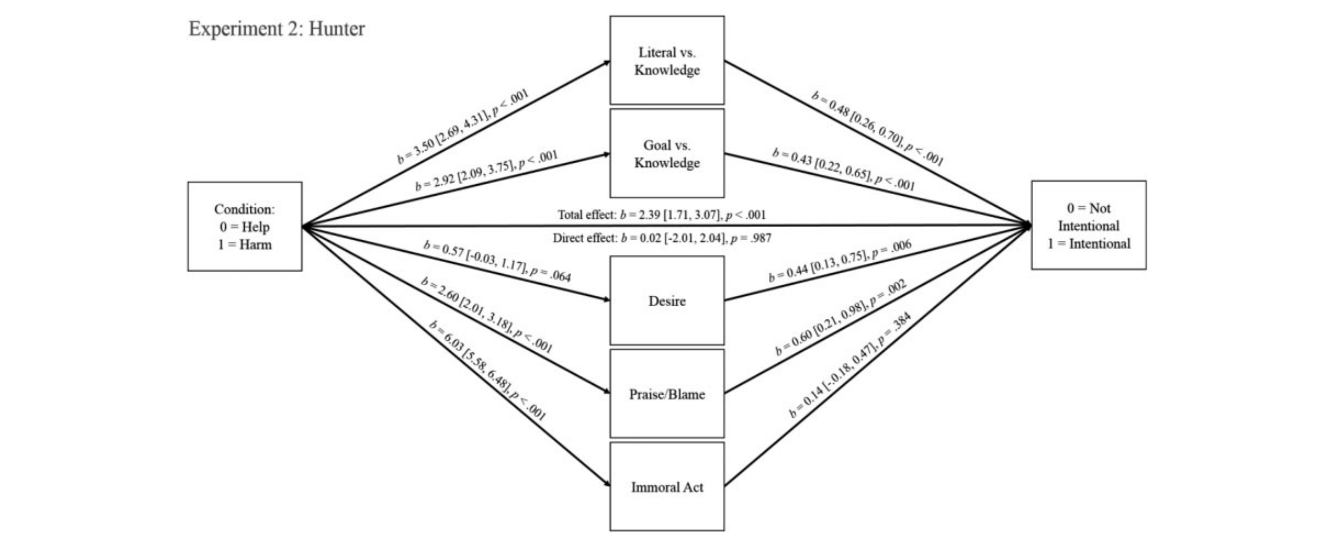
Experiment 2: Hunter
Condition: 0 = Help 1 = Harm
Literal vs. Knowledge
Goal vs. Knowledge
Desire
Praise/Blame
Immoral Act
0 = Not Intentional 1 = Intentional
b = 3.50 [2.69, 4.31], p < .001
b = 2.92 [2.09, 3.75], p < .001
Total effect: b = 2.39 [1.71, 3.07], p < .001
Direct effect: b = 0.02 [-2.01, 2.04], p = .987
b = 0.57 [-0.03, 1.17], p = .064
b = 2.60 [2.01, 3.18], p < .001
b = 6.03 [5.58, 6.48], p < .001
b = 0.48 [0.26, 0.70], p < .001
b = 0.43 [0.22, 0.65], p < .001
b = 0.44 [0.13, 0.75], p = .006
b = 0.60 [0.21, 0.98], p = .002
b = 0.14 [-0.18, 0.47], p = .384

Figure 5. Unstandardized path coefficients with 95% confidence intervals for mediation model in Experiment 2.

defaults), and what most people's attitudes would have been (i.e., general defaults). A similar approach was taken by Cova et al. (2016), who found that normative (i.e., personal) defaults mediated intentionality responses. Additional questions then asked about interpretations of the intentionality question, including whether participants believed the question was meant literally or was asking about the agent's foreknowledge, goals, or deservingness of praise/blame. We hypothesized that the intentionality question would again be defined in different ways across conditions, particularly regarding foreknowledge, and that condition-based differences in this latter variable would mediate intentionality responses beyond the effects of other variables.

Method

There were 111 female and 89 male participants (1 did not disclose; $M_{\text{age}} = 34.47$, $SD = 10.31$; no ideology question was included in Experiment 3 or 4). After reading help or harm versions of the chairman vignette and answering the same intentionality question used in Experiments 1 and 2, participants indicated the chairman's attitude ("What do you think the chairman's attitude was toward HH the environment?"), their personal default attitude ("What do you think the chairman's attitude should have been toward HH the environment?"), and their beliefs about the general-default attitude ("In a similar situation, what type of attitude would most people have toward HH the environment?") (–4

Table 3. Means, *SD*, *t*, *p*, *d*, and 95% CI for Experiment 3.

Dependent Variable	Help *M* (*SD*)	Harm *M* (*SD*)	*t*(199)	*p*	*d*	95% CI $M_{\text{Difference}}$
Perceived attitude	–0.54 (1.55)	0.92 (1.67)	6.45	<.001	0.90	[1.02, 1.91]
Personal default	2.48 (2.00)	–2.69 (2.13)	17.75	<.001	2.50	[4.59, 5.74]
General default	2.08 (1.66)	–1.80 (1.95)	15.20	<.001	2.14	[3.38, 4.38]
Meaning—literal	1.44 (3.09)	0.72 (2.79)	1.53	.129	0.24	[–0.10, 1.53]
Meaning—knowledge	–1.61 (2.76)	2.56 (2.15)	11.96	<.001	1.68	[3.49, 4.86]
Meaning—reason	–0.70 (3.22)	–1.35 (2.78)	1.72	.086	0.22	[–0.19, 1.48]
Meaning—praise/blame	–2.18 (2.46)	2.07 (2.49)	12.19	<.001	1.72	[3.56, 4.94]

Note. t, d, and CI are reported as positive values. *SD* = standard deviation; CI = confidence interval.

= strongly opposed to it, 4 = strongly in favor of it). For use in mediation tests, we respectively subtracted personal and general defaults from perceptions of the chairman's attitudes (i.e., "relative defaults"). Next, participants rated their agreement with four statements (presented in individualized random orders) (–4 = totally disagree, 4 = totally agree). Literal: "The question was literally asking whether the chairman's intentional action was to HH the environment." Knowledge: "The question was asking whether the chairman knew the environment would be HH when he started the program." Praise/blame: "The question was asking whether the chairman deserved praise/blame for the environment being HH." Reason: "The question was asking whether the chairman started the program to HH the environment."

Results

Intentionality responses. Very few people responded that the chairman intentionally helped the environment (6/101). Almost all participants responded that the chairman intentionally harmed the environment (95/100), $\chi^2(1) = 159.43$, $p < .001$, $\varphi = .89$.

Mean differences. Descriptive statistics, effect sizes, and the results of inferential tests are provided in Table 3. Harm (vs. help) condition participants thought the chairman's attitude was more pro-outcome, his attitude should be more con-outcome, and most people's attitudes would be more con-outcome. Harm (vs. help) condition participants also indicated that the intentionality question should be interpreted (a) descriptively less about the chairman's intentional action, (b) descriptively less about the chairman's reasons for acting, (c) significantly more as asking whether the chairman started the program, knowing the effect this would have on the environment, and (d) significantly more as asking whether he deserves blame (vs. praise) for the side effect. All means significantly differed from scale midpoints in both conditions, *p*s $\leq$ .001 (for "reason" in the help condition, $p = .030$; for "literal" in the harm condition, $p = .011$).

Repeated-measures *t* tests showed that in the help condition, the chairman's attitudes were significantly more con-helping than personal, $t(100) = –13.94$, $p < .001$, $d = 1.69$, 95% CI = [–3.45, –2.59], and general defaults, $t(100) = 13.29$, $p < .001$, $d = 1.63$, 95% CI = [–3.02, –2.23]. Personal defaults were significantly more pro-helping than general defaults, $t(100) = 2.45$, $p = .016$, $d = 0.22$, 95% CI = [0.08, 0.72]. In the harm condition, the chairman's attitudes were more pro-harming than personal, $t(99) = 14.45$, $p < .001$, $d = 1.89$, 95% CI = [3.11, 4.11], and general, $t(99) = 11.39$, $p < .001$, $d = 1.50$, 95% CI = [2.25, 3.19], defaults. General defaults were significantly less con-harming than personal defaults, $t(99) = –4.97$, $p < .001$, $d = –0.44$, 95% CI = [–1.25, –0.53].

Frequencies of knowledge definition responses. Because our primary interest is in examining whether people reinterpret questions about intentional harming as asking whether the agent intentionally did something, knowing it would lead to harm, we focus on responses to that question here (see Figure 6). Consistent with Experiments 1 and 2, more than half of help condition participants strongly disagreed that the intentionality question was asking whether the chairman knew the environment would be helped when he started the program. Strikingly, almost two thirds of harm condition participants strongly agreed that the question should be defined this way.

Mediation. Similar to Experiments 1 and 2, a simultaneous mediation test examined whether the effects of condition were carried to intentionality responses through any of the putative mediators (Figure 7). The indirect effect of condition on intentionality was significant through only one variable: belief that the intentionality question was asking whether the chairman started the program, knowing it would help/harm the environment, $ab = 0.99$, 95% CI = [0.16, 1.93]. Indirect effects through a literal interpretation of the question ($ab = 0.03$, 95% CI = [–0.12, 0.49]), a reason-based interpretation ($ab = 0.04$, 95% CI = [–0.09, 0.37]), and a praise/blame interpretation ($ab = –0.21$, 95% CI = [–1.63, 1.13]) were not significant, as 95% CIs contained zero. Similarly, the relation of attitudes to personal ($ab = 1.22$, 95% CI = [–1.26, 3.9]) and general defaults ($ab = 0.65$, 95% CI = [–1.29, 2.78]) were not significant mediators. The

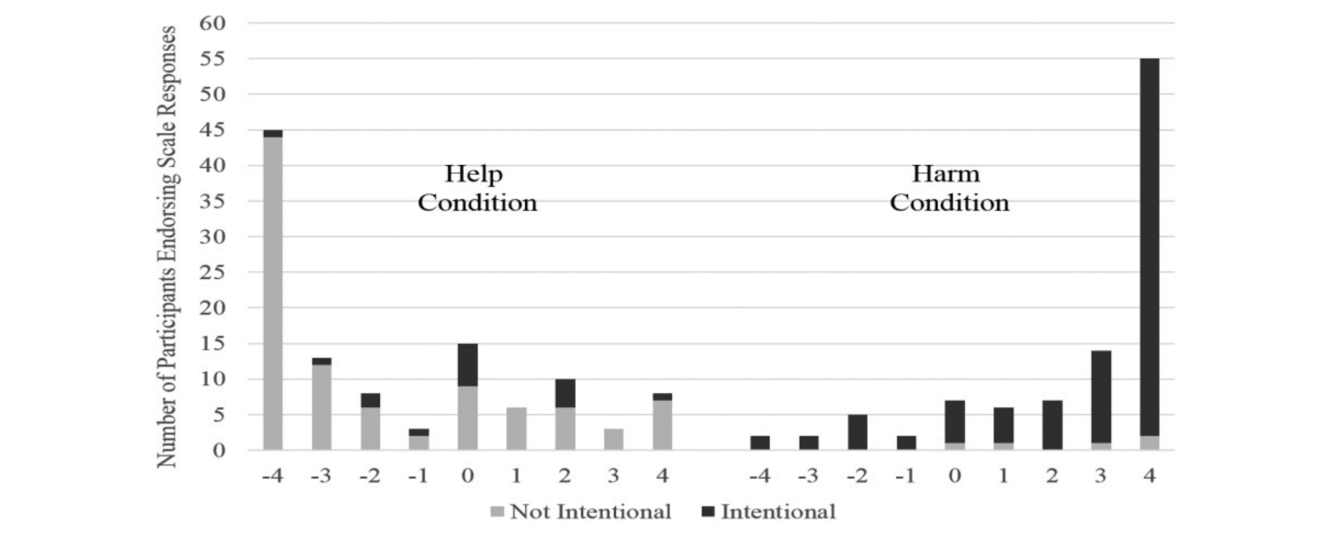

Figure 6. Frequencies of labeling helping or harming as intentional as a function of responses to the knowledge question in Experiment 3.

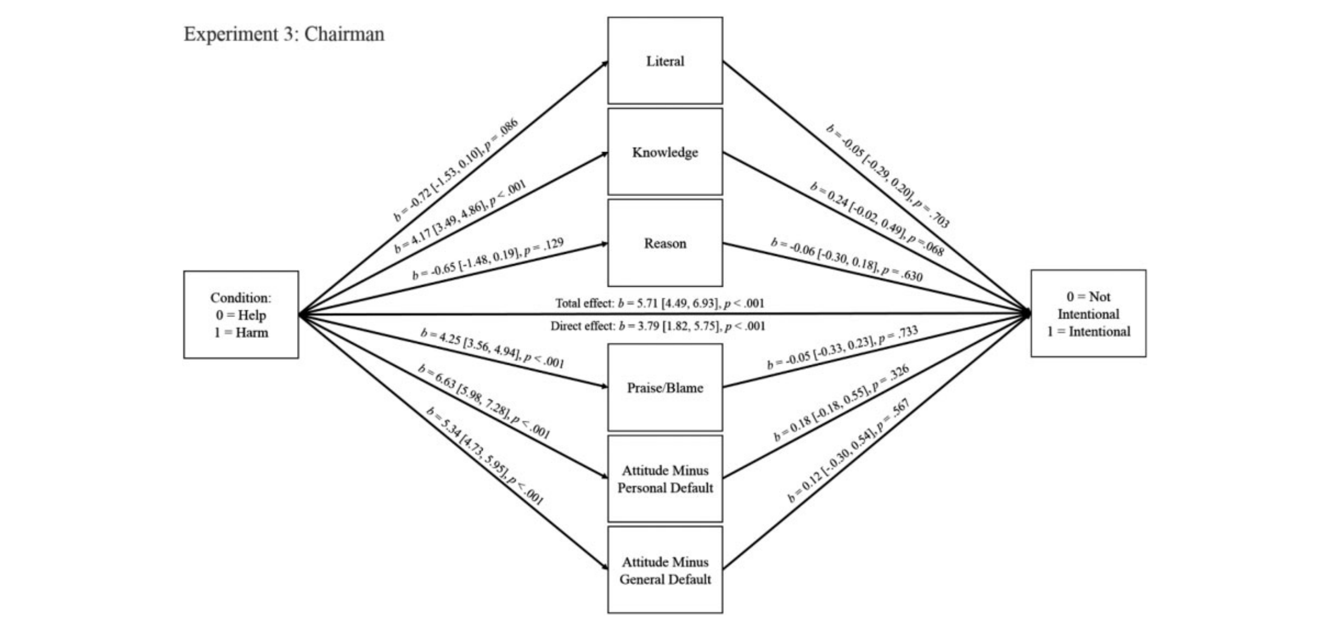
Experiment 3: Chairman
Condition: 0 = Help 1 = Harm
Literal
Knowledge
Reason
Praise/Blame
Attitude Minus Personal Default
Attitude Minus General Default
0 = Not Intentional 1 = Intentional
b = -0.72 [-1.53, 0.10], p = .086
b = 4.17 [3.49, 4.86], p < .001
b = -0.65 [-1.48, 0.19], p = .129
Total effect: b = 5.71 [4.49, 6.93], p < .001
Direct effect: b = 3.79 [1.82, 5.75], p < .001
b = 4.25 [3.56, 4.94], p < .001
b = 6.63 [5.98, 7.28], p < .001
b = 5.34 [4.73, 5.95], p < .001
b = -0.05 [-0.29, 0.20], p = .703
b = 0.24 [-0.02, 0.49], p =.068
b = -0.06 [-0.30, 0.18], p = .630
b = -0.05 [-0.33, 0.23], p = .733
b = 0.18 [-0.18, 0.55], p = .326
b = 0.12 [-0.30, 0.54], p = .567

Figure 7. Unstandardized path coefficients with 95% confidence intervals for mediation model in Experiment 3.

direct effect of condition on intentionality remained significant, $c' = 3.79$, 95% CI = [1.82, 5.75]. Because of a high correlation between personal and normative defaults ($r = .84$), additional models tested the relative defaults in separate models, and also tested the individual indirect effects of perceived attitudes, personal defaults, and general defaults (i.e., rather than relative defaults) in separate multiple-mediator models. In each of these analyses, the only significant indirect effect was through defining the intentionality question (or not) in terms of acting with foreknowledge. Three exploratory analyses also examined single-mediator models containing only perceived attitudes, personal defaults, and general defaults as mediators. None of these variables significantly mediated condition on intentionality responses.

Discussion

As in Experiments 1 and 2, participants in Experiment 3 defined the intentionality question in different ways across conditions. Harm (vs. help) condition participants defined it somewhat less as asking whether the chairman started the program to harm, and significantly more as asking whether the chairman deserved blame (vs. praise in the help condition). Harm (vs. help) condition participants were also somewhat less certain that the intentionality question was asking whether the chairman's intentional action was literally to harm the environment. However, harm condition participants appeared fairly certain that it *was* asking whether he (intentionally) *started the program, knowing* the environment would be harmed. Although subtle, this difference is important, as the foreknowledge interpretation matches the stated facts of the story. In the help condition, most participants rejected this redefinition. Follow-up repeated-measures *t*-tests confirmed that harm condition participants more strongly endorsed this definition than the literal (intentional action) question, with the reverse true in the help condition (*p*s < .001). The extent of agreement with the foreknowledge interpretation of the intentionality question was also the only variable that significantly mediated responses to the intentionality question.

Notably, attitudes relative to defaults, attitudes on their own, and defaults on their own (even when tested in single-mediator models) consistently failed to mediate intentionality responses. However, this runs counter to an earlier finding (Cova et al., 2016), where in two experiments, personal ("normative") defaults did mediate intentionality responses. A number of methodological differences between the current work and Cova et al. (2016) might help explain why the effect failed to reproduce here. Specifically, in Cova et al. (2016), (a) a continuous measure of intentionality was used, (b) several different experimental conditions were combined for use in each mediation test, with one experiment manipulating whether the chairman was joyful versus regretful about helping/harming and a second experiment manipulating whether the chairman was regretful versus indifferent about harming or joyful versus indifferent about helping, (c) the helping/harming was concretely described (i.e., the profit-increasing programs released organic fertilizers or toxic chemicals that helped a forest grow and bloom or destroyed it) and the outcomes may have varied in perceived impact, and (d) the personal default measure was different than was used here. Also, no questions about definitions of intentionality questions were included and simultaneously examined. Any of these differences might help explain the difference in findings. Of course, the lack of significant mediation through defaults or other variables here does not allow strong conclusions to be drawn that indirect effects through these variables are zero in the population. Yet, the clear and reproducible differences across conditions in people's explicit beliefs about what the intentionality question was asking does cast doubt on whether responses to the intentionality question should be unreservedly interpreted as reflecting differences in their intuitions about intentional action.

Other findings regarding the chairman's attitudes and the two defaults were also informative. Most people in both conditions (75% in the help condition, 65% in the harm condition) believed the chairman was neither in favor of nor opposed to helping/harming. This corresponds with information provided in the story itself and suggests that at least for side-effect cases where the agents' attitudes are explicitly described as indifferent, perceptions of these attitudes might be treated as a near-constant. If so, what theoretically matters most in these cases might be perceived locations of defaults. Regarding the relation of the chairman's perceived attitude to defaults (Knobe, 2010)—particularly personal defaults—it would be virtually impossible for participants to ever view the chairman's con-attitudes as equally or more con than the default regarding harming the environment when using the standard chairman vignette. That is, 52% of participants in the harm condition endorsed the most extreme value on the 9-point personal-default scale ("strongly opposed"), and another 23% endorsed the next most extreme value. In the help condition, 60.4% of participants endorsed the most extreme or next most extreme value in the opposite direction ("strongly in favor"). Although general-default perceptions were less extreme, 38% and 40% of people, respectively, believed that most people would be in strongly in favor of helping and strongly against harming, even though one might realistically expect a company's main concern to be generating profits, not the environment (e.g., Laufer, 2003). Thus, even when perceived as opposed to harming or in favor of helping, the chairman could not be viewed as more con or pro than the highest possible values of the scales, making the attitude-relative-to-default hypothesis difficult to falsify. This is particularly true because even if people believed he was opposed to harming, he *did* start the program and the environment *was* harmed, which invites a question about how opposed he really was. A final experiment explored this question by describing the chairman's attitudes as strongly pro-environment in both conditions, such that participants would view him as more strongly in favor of helping and opposed to helping.

Experiment 4

In Experiments 1 to 3, there was little within-condition variability in intentionality responses; most participants labeled harming as intentional, but few labeled helping as intentional. Prior research has shown that when agents' attitudes toward side effects uphold rather than violate norms, labeling of helping as intentional can increase enough to fully attenuate the SEE (Laurent et al., 2019). Potentially, this shift in attitudes increases the likelihood that help condition participants will interpret the intentionality question as asking whether the agent acted to help, wanted to help, and foresaw the helpful outcome when acting. In harm conditions, although participants might believe the agent is strongly against harming, that he acts anyway should continue to result in the intentionality question mostly being interpreted as asking about an intentional action undertaken with foreknowledge of harm.

Experiment 4, using an adapted version of the chairman scenario (Laurent et al., 2019; see also Cova et al., 2012;

Table 4. Means, *SD*, *t*, *p*, *d*, and 95% CI for Experiment 4.

Dependent Variable	Help *M* (*SD*)	Harm *M* (*SD*)	*t*(302)	*p*	*d*	95% CI $M_{\text{Difference}}$
Perceived attitude	2.43 (1.75)	–0.59 (2.15)	13.39	<.001	1.54	[2.57, 3.46]
Personal default	2.41 (1.59)	−1.22 (2.48)	15.10	<.001	1.74	[3.16, 4.10]
General default	1.41 (1.85)	–0.16 (2.28)	6.55	<.001	0.76	[1.10, 2.04]
Goal (profit vs. HH)	−1.23 (2.97)	–2.50 (2.51)	4.03	<.001	0.46	[0.65, 1.89]
Meaning—literal	1.49 (2.20)	–0.28 (2.94)	5.91	<.001	0.68	[1.18, 2.36]
Meaning—knowledge	0.99 (2.59)	1.89 (2.31)	3.22	.001	0.37	[0.35, 1.46]
Meaning—reason	1.26 (2.44)	–0.74 (3.01)	6.37	<.001	0.73	[1.39, 2.62]
Meaning—praise/blame	–0.57 (2.63)	1.12 (2.48)	5.77	<.001	0.66	[1.11, 2.27]
Meaning—want	1.36 (2.42)	–0.66 (2.74)	6.79	<.001	0.78	[2.43, 2.60]

Note. t, d, and CI are reported as positive values. CI = confidence interval; HH = help/harm.

Shepherd, 2012), tested this hypothesis. In it, the chairman's company is badly in need of profits, and the chairman in both conditions is described as having a strong record of advocating for environmental protection. In the help version, he is happy the environment will be helped but acknowledges that the company's main goal is to increase profits. In the harm version, he is upset about harming and describes the decision as difficult, but indicates that because of the company's financial troubles, they must start the program.

Method

Participants, procedure, and measures. Three-hundred four people participated (152 females, 149 males, three undisclosed; $M_{\text{age}} = 35.57$, $SD = 11.74$). After reading an adapted help or harm version of the chairman vignette, participants responded to the same intentionality question from Experiments 1 to 3. The same three questions from Experiment 3 about the chairman's attitudes, personal defaults, and general defaults followed. For mediation tests, variables capturing attitudes minus personal and general defaults were again created. One question asked about the chairman's main goal: "In starting the program, what was the chairman's main goal? His goal was . . . " (–4 = increasing profits, 4 = HH the environment). This was included to confirm that helping/harming (primarily helping) was viewed as a side effect, rather than the goal motivating the chairman's action. Following this, five items, presented in individualized random orders and using the same 9-point agreement scale as in Experiment 3, asked about the perceived meaning of the intentionality question. Literal, knowledge, reason, and praise/blame items were the same as in Experiment 3. A fifth item asked whether the question was about the chairman's desire to HH: "The question was asking whether the chairman wanted to HH the environment."

Results

Intentionality responses. A majority of people (101/148) in the help condition responded that the chairman intentionally helped the environment. As expected, a larger majority in the harm condition (129/156) responded that the chairman intentionally harmed the environment, $\chi^2(1) = 8.61$, $p = .003$, $\varphi = .17$.

Mean differences. Descriptive statistics, effect sizes, and the results of inferential tests are provided in Table 4. Help (vs. harm) condition participants thought the chairman's attitude was more pro-outcome, his attitude should be more pro-outcome, but that most people's attitudes would be *less* pro-outcome. Help (vs. harm) condition participants also thought the chairman's goal was less profit-oriented. Importantly, responses to this latter question were significantly below scale midpoints in both conditions (*p*s < .001), indicating that people generally believed his main goal was more profit-oriented than side-effect-outcome–oriented. Harm (vs. help) condition participants thought the intentionality question should be interpreted less in terms of the chairman's (a) intentional action, (b) reasons for acting, and (c) desire for the side effect, and more as asking whether the chairman (d) intentionally acted, knowing the effect this would have on the environment, and (e) deserves blame for the side effect. All means significantly differed from scale midpoints in the help condition, $ps \leq .009$. In the harm condition, all questions but the general default ($p = .382$) and literal interpretation of the intentionality question ($p = .244$) significantly differed from scale midpoints, $ps \leq .003$.

Repeated-measures *t* tests showed that in the help condition, the chairman's attitudes did not significantly differ from personal defaults, $t(147) = 0.09$, $p = .932$, $d = 0.01$, 95% CI = [–0.30, 0.33], and were *more* pro-helping than general defaults, $t(147) = 5.35$, $p < .001$, $d = 0.58$, 95% CI = [0.64, 1.40]. Personal defaults were significantly more pro-helping than general defaults, $t(147) = 7.59$, $p < .001$, $d = 0.58$, 95% CI = [0.74, 1.27]. In the harm condition, the chairman's attitudes were more pro-harming than personal defaults, $t(155) = 3.31$, $p = .001$, $d = 0.27$, 95% CI = [0.25, 1.00], but significantly *less* pro-harming than general defaults, $t(155) = –2.00$, $p = .047$, $d = 0.19$, 95% CI =

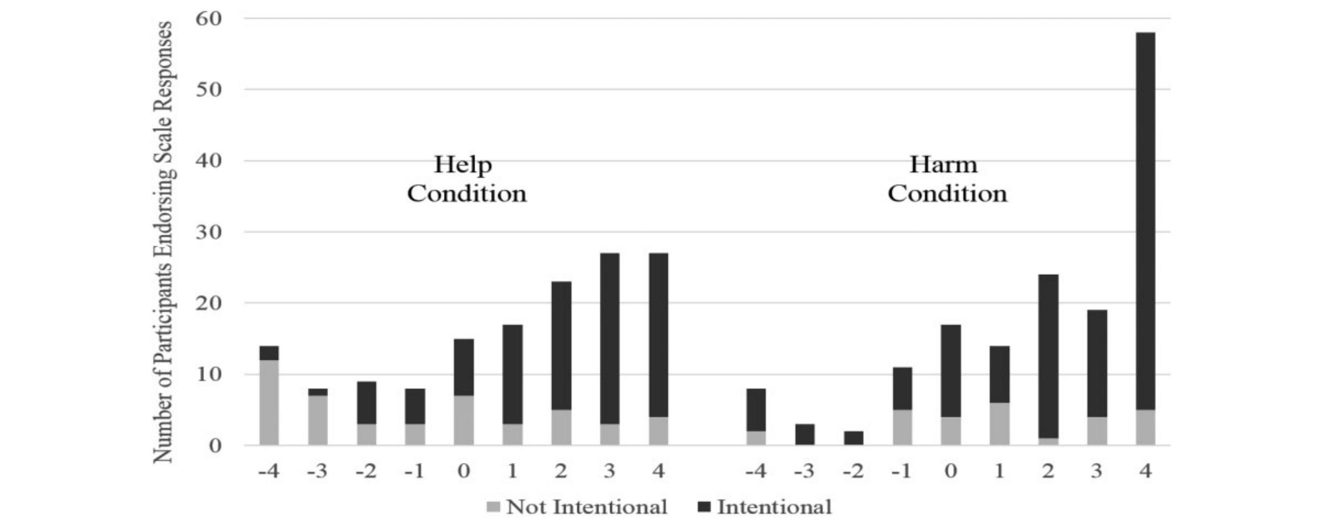

Figure 8. Frequencies of labeling helping or harming as intentional as a function of responses to the knowledge question in Experiment 4.

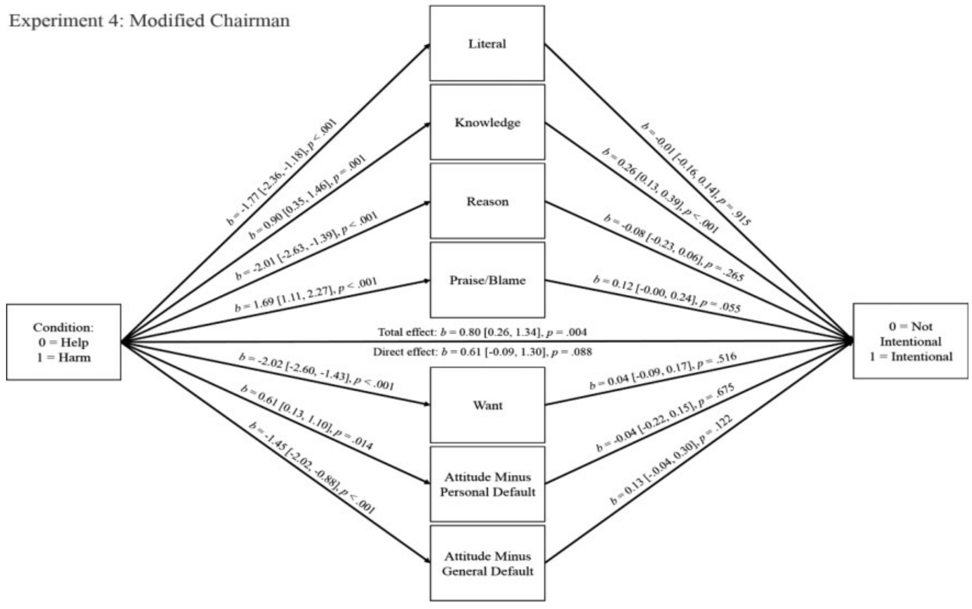
Experiment 4: Modified Chairman
Condition:
0 = Help
1 = Harm
Literal
Knowledge
Reason
Praise/Blame
Want
Attitude Minus Personal Default
Attitude Minus General Default
0 = Not Intentional
1 = Intentional
b = -1.77 [-2.36, -1.18], p < .001
b = 0.90 [0.35, 1.46], p = .001
b = -2.01 [-2.63, -1.39], p < .001
b = 1.69 [1.11, 2.27], p < .001
Total effect: b = 0.80 [0.26, 1.34], p = .004
Direct effect: b = 0.61 [-0.09, 1.30], p = .088
b = -2.02 [-2.60, -1.43], p < .001
b = 0.61 [0.13, 1.10], p = .014
b = -1.45 [-2.02, -0.88], p < .001
b = -0.01 [-0.16, 0.14], p = .915
b = 0.26 [0.13, 0.39], p < .001
b = -0.08 [-0.23, 0.06], p = .265
b = 0.12 [-0.00, 0.24], p = .055
b = 0.04 [-0.09, 0.17], p = .516
b = -0.04 [-0.22, 0.15], p = .675
b = 0.13 [-0.04, 0.30], p = .122

Figure 9. Unstandardized path coefficients with 95% confidence intervals for mediation model in Experiment 4.

[–0.85, –0.01]. General defaults were significantly less pro-harming than personal defaults, $t(155) = –5.67$, $p < .001$, $d = –0.44$, 95% CI = [–1.42, –0.69]. Notably, although the chairman's attitudes were perceived as more opposed to harming than the general default, most participants (129/156, or 83%) still labeled harming as intentional. In addition, although the chairman's attitudes did not differ from personal defaults and were *more* pro-helping than general defaults, about a third of participants responded that he did *not* intentionally help, suggesting that the relation of attitudes to defaults is not sufficient to fully explain the pattern of intentionality responses.

Frequencies of knowledge definition responses. Unlike Experiment 3, many participants in the help condition agreed that the question was asking whether the chairman started the program, knowing the environment would be helped. In the harm condition, many people continued to strongly agree with this definition. Figure 8 provides frequencies of responses to the intentionality question as a function of how people responded to the knowledge definition question.

Mediation. A simultaneous mediation test was again performed (Figure 9). The indirect effect of condition on intentionality responses was again significant through knowledge,

ab = 0.24, 95% CI = [0.07, 0.48]. As in Experiment 3, no other indirect effects were significant: attitude minus personal default (ab = –0.02, 95% CI = [–0.19, 0.09]); attitude minus general default (ab = –0.19, 95% CI = [–0.53, 0.07]); literal interpretation (ab = 0.01, 95% CI = [–0.28, 0.35]); reason interpretation (ab = 0.17, 95% CI = [–0.14, 0.57]); praise/blame interpretation (ab = 0.20, 95% CI = [–0.02, 0.47]); desire for outcome interpretation (ab = –0.09, 95% CI = [–0.39, 0.20]). The direct effect of condition was not significant, c' = 0.61, 95% CI = [–0.09, 1.30]. Also, as in Experiment 3, multiple mediation models were examined testing relative defaults separately, attitudes and each default separately, and attitudes and each default in single-variable-mediator models. Indirect effects were not significant for any of these tests (ps > .05), with one exception: When tested in isolation, perceived attitudes mediated intentionality responses (p < .05) such that the chairman's greater pro-attitudes were associated with *less* labeling of the side effect as intentional.

Discussion

Experiment 4 used a version of the chairman vignette where profits were badly needed to avoid financial catastrophe, starting a new program would increase profits but also help/harm the environment, and the chairman was an environmental advocate. Comparing this descriptively with Experiments 1 to 3, a much larger percentage of help condition participants labeled helping as intentional. Importantly, participants in both conditions indicated that the chairman's action was primarily directed at increasing profits (i.e., not helping/harming), which suggests that effects on the environment remained a side effect and were not viewed as the chairman's primary goal. This change in the chairman story was also associated with an increased belief in the help condition that the intentionality question was asking whether the chairman started the program, knowing it would lead to the environment being helped. This is the same interpretation that participants in harm conditions strongly and consistently endorsed. As in Experiment 3 (and similar to findings in Experiments 1 to 2), this variable alone significantly mediated intentionality responses.

General Discussion

In four experiments (and two replications), the following question was explored: "Do people define questions asking about the intentionality of side-effect helping versus harming differently?" A secondary question asked whether, if so, would these differences help explain the SEE? The short answer to both is, yes, we think so.

Using the standard chairman vignette (Knobe, 2003), a different replication vignette (adapted from Nadelhoffer, 2006), and an adapted chairman vignette, we found systematic differences in how people interpreted a question asking about the intentionality of side effects. Experiments 1 and 2 found that most participants in help conditions believed the question was literally asking whether the agents' intentional action was to help (or in replications, whether the agents intentionally helped). In harm conditions, most participants believed the question was asking whether the agents intentionally acted, knowing that harm would result. Similarly, help condition participants indicated on average that the chairman's goals were more important than his foreknowledge to their responses. Harm condition participants indicated the opposite. In Experiments 3 and 4, relative to help conditions, participants in harm conditions defined the question less in terms of the chairman's intentional action and more as asking whether the chairman started a program, knowing it would harm the environment, more as asking whether the chairman deserves blame (vs. praise), and less as asking about whether the chairman acted to bring about the side effect. In Experiment 4, they also defined it less as asking whether the chairman wanted to harm the environment.

These reliable differences in how the intentionality question is interpreted highlight the difficulty in firmly concluding that differences in side-effect outcome valence, through whatever process (including normative defaults; Cova et al., 2016), fundamentally impact people's intuitions about intentional action, even if they impact *responses* to questions about the intentionality of helping/harming. That is, if the majority of people's responses to the ostensibly same intentionality question are based on substantially different understanding of the question's meaning, comparing responses across conditions and drawing conclusions based on a straightforward interpretation of the question is problematic. This is especially true when drawing conclusions about affirmative responses in harm cases, where majorities of participants explicitly indicate that they do not understand the question to be meant literally, as written (e.g., replications of Experiments 1 and 2; see also Laurent, Clark, & Schweitzer, 2015; Laurent et al., 2019).

In these cases, people appear to reconstruct the intentionality question in two primary ways. The first—at least in cases where blame is a relevant concept (cf. Knobe & Mendlow, 2004)—is to interpret it as asking whether the chairman deserves blame for his behavior (e.g., Alicke & Rose, 2010). This makes sense because harming agents' behavior appears quite immoral and blameworthy. Even if agents' reasons for acting are not specifically to harm, it seems wrong for them to intentionally do things they know will lead to harm, particularly when they appear unconcerned about this undesirable outcome. In contrast, indifference about the beneficial side effects of one's behavior seems especially undeserving of praise. Why glorify those who help when they are obviously indifferent to helping or perhaps even hostile toward it? The obviousness of this may underlie why help condition participants are less likely to interpret intentionality questions as asking whether helping agents deserve praise.

Second, questions about intentionality of harm may focus people on two distinct elements presented in the vignette: the agent's *intentional action* (e.g., starting a profit-increasing program) and the harmful secondary outcome he knows this goal-directed action will cause. Because the concept of intentionality is most frequently applied to actions rather than consequences of actions (Laurent, Clark, & Schweitzer, 2015), reframing the question as asking about an intentional action undertaken with foreknowledge of harm has advantages. It allows consideration of key elements from the story and is responsive to what people may feel is at the heart of the question: "Did the chairman act intentionally, knowing this would lead to harm?" Notably, responses to questions capturing this idea significantly mediated intentionality responses in each experiment presented here, whereas other variables tested failed to consistently do so. Although statistical mediation is not determinative, this replication across four experiments and two replications is suggestive. Of course, we acknowledge that in research using different methodological approaches, other mediators such as normative defaults have significantly mediated intentionality responses (Cova et al., 2016). This keeps open the possibility that perceived attitudes, defaults, or attitudes relative to defaults (Knobe, 2010) might help explain participants' intentionality responses in some cases, even if they failed to do so here.

However, even if moral judgments or attitudes relative to defaults had significantly mediated the effects of condition on intentionality here, interpretations would require caution. That is, this would not necessarily provide strong support for the view that intuitions about intentionality differ as a function of side-effect outcome valence. Instead, it might help explain why people sometimes understand questions about the intentionality of side effects in different ways. Sometimes they are seen as asking what they appear to ask; at other times, they seem to be asking something different.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Sean M. Laurent https://orcid.org/0000-0003-0130-7867
Brandon J. Reich https://orcid.org/0000-0002-2977-4339

Notes

[1] Original anchors were from 1 to 9. For descriptive purposes, we recoded these throughout to center on scale midpoints. In Experiment 1, praise/blame was presented on a 10-point scale because of a programming error.

[2] A question asking about the relative importance of the agents' action versus the side-effect outcome for their intentionality responses was also included. Because it is not relevant here, results of this variable are not discussed further.

[3] We thank an anonymous reviewer for this suggestion.

[4] We also conducted a similar replication of Experiment 2 ($N = 199$). Results of this replication were fully consistent with those reported here and are also described in the Supplemental Appendix of this article.

Supplemental Material

Supplemental material for this article is available online.

References

Adams, F., & Steadman, A. (2004). Intentional action and moral considerations: Still pragmatic. *Analysis*, *64*(3), 268-276. https://doi.org/10.1093/analys/64.3.268

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. 17-33). Ashgate Publishing.

Alicke, M. (2008). Blaming badly. *Journal of Cognition and Culture*, *8*(1), 179-186. https://doi.org/10.1163/156770908X289279

Alicke, M., & Rose, D. (2010). Culpable control or moral concepts? *Behavioral and Brain Sciences*, *33*(4), 330-331. https://doi.org/10.1017/S0140525X10001664

Cova, F., Dupoux, E., & Jacob, P. (2012). On doing things intentionally. *Mind & Language*, *27*(4), 378-409. https://doi.org/10.1111/j.1468-0017.2012.01449.x

Cova, F., Lantian, A., & Boudesseul, J. (2016). Can the Knobe effect be explained away? Methodological controversies in the study of the relationship between intentionality and morality. *Personality and Social Psychology Bulletin*, *42*(10), 1295-1308. https://doi.org/10.1177/0146167216656356

Cova, F., & Naar, H. (2012). Side-effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality. *Philosophical Psychology*, *25*(6), 837-854. https://doi.org/10.1080/09515089.2011.622363

Cushman, F., & Mele, A. (2008). Intentional action: Two-and-a-half folk concepts? In J. Knobe & S. Nichols (Eds.), *Experimental philosophy* (pp. 171-188). Oxford University Press.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, *7*(4), 324-336. https://doi.org/10.1207/S15327957PSPR0704_05

Fitzgerald, P. J., & Williams, G. (1962). Carelessness, indifference, and recklessness: Two replies. *The Modern Law Review*, *25*(1), 49-58. https://doi.org/10.1111/j.1468-2230.1962.tb00678.x

Garner, B. A. (2014). *Black's law dictionary* (10th ed.). West Publishing.

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, *36*(12), 1635-1647. https://doi.org/10.1177/0146167210386733

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLOS ONE*, *14*(3), Article e0213544. https://doi.org/10.1371/journal.pone.0213544

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557-559. https://doi.org/10.1038/nature06288

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*(3), 521-537. https://doi.org/10.1037/a0013779

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190-194. https://doi.org/10.1111/1467-8284.00419

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315-329. https://doi.org/10.1017/S0140525X10000907

Knobe, J., & Mendlow, G. S. (2004). The good, the bad and the blameworthy: Understanding the role of evaluative reasoning in folk psychology. *Journal of Theoretical and Philosophical Psychology*, *24*(2), 252-258. https://doi.org/10.1037/h0091246

Lanteri, A. (2012). Three-and-a-half folk concepts of intentional action. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *158*, 17-30. https://doi.org/10.2307/41406995

Laufer, W. S. (2003). Social accountability and corporate greenwashing. *Journal of Business Ethics*, *43*(3), 253-261. https://doi.org/10.1023/A:1022962719299

Laurent, S. M., Clark, B. A. M., & Schweitzer, K. A. (2015). Why side-effect outcomes do not affect intuitions about intentional actions: Properly shifting the focus from intentional outcomes back to intentional actions. *Journal of Personality and Social Psychology*, *108*(1), 18-36. https://doi.org/10.1037/pspa0000011

Laurent, S. M., Nuñez, N. L., & Schweitzer, K. A. (2015). The influence of desire and knowledge on perception of each other and related mental states, and different mechanisms for blame. *Journal of Experimental Social Psychology*, *60*, 27-38. https://doi.org/10.1016/j.jesp.2015.04.009

Laurent, S. M., Nuñez, N. L., & Schweitzer, K. A. (2016). Unintended, but still blameworthy: The roles of awareness, desire, and anger in negligence, restitution, and punishment. *Cognition and Emotion*, *30*(7), 1271-1288. https://doi.org/10.1080/02699931.2015.1058242

Laurent, S. M., Reich, B. J., & Skorinko, J. L. M. (2019). Reconstructing the side-effect effect: A new way of understanding how moral considerations drive intentionality asymmetries. *Journal of Experimental Psychology: General*, *148*, 1747-1766. https://doi.org/10.1037/xge0000554

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language*, *23*(2), 165-189. https://doi.org/10.1111/j.1468-0017.2007.00336.x

Malle, B. F. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture*, *6*(1), 87-112. https://doi.org/10.1163/156853706776931358

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186. https://doi.org/10.1080/1047840X.2014.877340

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, *102*(4), 661-684. https://doi.org/10.1037/a0026790

Malle, B. F., & Knobe, J. (1997a). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101-121. https://doi.org/10.1006/JESP.1996.1314

Malle, B. F., & Knobe, J. (1997b). Which behaviors do people explain? A basic actor–observer asymmetry. *Journal of Personality and Social Psychology*, *72*(2), 288-304. https://doi.org/10.1037/0022-3514.72.2.288

Malle, B. F., Moses, L. J., & Baldwin, D. A. (2001). *Intentions and intentionality: Foundations of social cognition*. MIT Press.

Mele, A. R., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, *31*(1), 184-201. https://doi.org/10.1111/j.1475-4975.2007.00147.x

Nadelhoffer, T. (2006). Desire, foresight, intentions, and intentional actions: Probing folk intuitions. *Journal of Cognition and Culture*, *6*(1), 133-157. https://doi.org/10.1163/156853706776931259

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe Effect revisited. *Mind & Language*, *22*(4), 346-365. https://doi.org/10.1111/j.1468-0017.2007.00312.x

Nuñez, N., Laurent, S., & Gray, J. M. (2014). Is negligence a first cousin to intentionality? Lay conceptions of negligence and its relationship to intentionality. *Applied Cognitive Psychology*, *28*(1), 55-65. https://doi.org/10.1002/acp.2957

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, *24*(5), 586-604. https://doi.org/10.1111/j.1468-0017.2009.01375.x

Rosset, E. (2008). It's no accident: Our bias for intentional explanations. *Cognition*, *108*(3), 771-780. https://doi.org/10.1016/j.cognition.2008.07.001

Royzman, E., & Hagan, J. P. (2017). The shadow and the tree: Inference and transformation of cognitive content in psychology of moral judgment. In J.-F. Bonnefon & B. Trémolière (Eds.), *Moral inferences: Current issues in thinking and reasoning* (pp. 56-74). Routledge.

Shepherd, J. (2012). Action, attitude, and the Knobe effect: Another asymmetry. *Review of Philosophy and Psychology*, *3*(2), 171-185. https://doi.org/10.1007/s13164-011-0079-7

Stark, F. (2016). *Culpable carelessness*. https://doi.org/10.1017/CBO9781139855945

Vaish, A., Carpenter, M., & Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. *Child Development*, *81*(6), 1661-1669. https://doi.org/10.1111/j.1467-8624.2010.01500.x