Journal of Experimental Psychology: General

Reconstructing the Side-Effect Effect: A New Way of Understanding How Moral Considerations Drive Intentionality Asymmetries

Sean M. Laurent, Brandon J. Reich, and Jeanine L. M. Skorinko

Online First Publication, January 17, 2019. http://dx.doi.org/10.1037/xge0000554

CITATION

Laurent, S. M., Reich, B. J., & Skorinko, J. L. M. (2019, January 17). Reconstructing the Side-Effect Effect: A New Way of Understanding How Moral Considerations Drive Intentionality Asymmetries. *Journal of Experimental Psychology: General*. Advance online publication. http://dx.doi.org/10.1037/xge0000554

American Psychological Association

Journal of Experimental Psychology: General

© 2019 American Psychological Association
0096-3445/19/$12.00

http://dx.doi.org/10.1037/xge0000554

Reconstructing the Side-Effect Effect: A New Way of Understanding How Moral Considerations Drive Intentionality Asymmetries

Sean M. Laurent
University of Illinois Urbana–Champaign

Brandon J. Reich
University of Oregon

Jeanine L. M. Skorinko
Worcester Polytechnic Institute

People typically apply the concept of intentionality to actions directed at achieving desired outcomes. For example, a businessperson might intentionally start a program aimed at increasing company profits. However, if starting the program leads to a foreknown and harmful side effect (e.g., to the environment), the side effect is frequently labeled as intentional even though it was not specifically intended or desired. In contrast, positive side effects (e.g., helping the environment) are rarely labeled as intentional. One explanation of this *side-effect effect*—that harmful (but not helpful) side effects are labeled as intentional—is that moral considerations influence whether people view actions as intentional or not, implying that bad outcomes are perceived as more intentional than good outcomes. The present research, however, shows that people redefine questions about intentionality to focus on agents' foreknowledge in harming cases and on their lack of desire or intention in helpful cases, suggesting that the same intentionality question is being interpreted differently as a function of side effect valence. Consistent with this, removing foreknowledge lowers the frequency of labeling harming as intentional without affecting whether people label helping as intentional. Likewise, increasing agents' desire to help or avoid harming increases rates of labeling helping as intentional without affecting rates of labeling harming as intentional. In summary, divergent decisions to label side effects as intentional or not appear to reflect differences in the criteria people use to evaluate each case, resulting in different interpretations of what questions about intentionality are asking.

Keywords: intentionality, side-effect effect, foreknowledge, desire/goals, moral judgment

Supplemental materials: http://dx.doi.org/10.1037/xge0000554.supp

Imagine being presented with the following story, first described in Knobe (2003a):

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

A question is then posed to you: "Did the chairman intentionally harm the environment?" Most people respond that the chairman intentionally harmed. Now imagine the same story with one change: Instead of harming it, the new program helps the environment. Everything else is the same. The chairman still knows what will happen and still doesn't care. People are asked: "Did the chairman intentionally help the environment?" Most people respond no, he did not intentionally help.

This finding has been called the side-effect effect (or the Knobe effect), and it suggests that harmful outcomes may be seen as intentional even when agents do not particularly desire or specifically intend these outcomes to occur. However, this runs counter to models suggesting that desire and intention are necessary inputs to judging an action as intentional (e.g., Adams, 1986; Malle & Knobe, 1997). Practically, addressing questions about how people judge intentionality in these cases is important because intentionally caused harms are viewed as deserving of more punishment than unintended harms, such as those that arise from negligence or recklessness (e.g., Darley & Pittman, 2003; Robinson & Darley, 1995). Theoretically, the same questions are important because intentions and intentionality are foundational to social cognition (e.g., Gergely, Nádasdy, Csibra, & Bíró, 1995; Malle, Moses, & Baldwin, 2001) and have been previously assumed to play a key causal role in moral judgments starting in childhood (e.g., Hamlin, Wynn, & Bloom, 2007; Vaish, Carpenter, & Tomasello, 2010),

Sean M. Laurent, Department of Psychology, University of Illinois Urbana–Champaign; Brandon J. Reich, Department of Marketing, University of Oregon; Jeanine L. M. Skorinko, Department of Social Science & Policy Studies, Worcester Polytechnic Institute.

We thank Dov Cohen and Dolores Albarracín for their helpful comments on a draft of this article.

Correspondence concerning this article should be addressed to Sean M. Laurent, Department of Psychology, University of Illinois Urbana–Champaign, 603 East Daniel Street, Champaign, IL 61820. E-mail: slauren@illinois.edu or seanmlaurent@gmail.com

rather than the reverse. Ultimately, given the importance of evaluating intentionality for everyday human interactions (e.g., Malle & Hodges, 2005; Reeder, 2009; Rosset, 2008; Rosset & Rottman, 2014), one imperative question is: Why does this effect emerge?

One proposed answer is that moral considerations underpin people's reasoning about intentional action, leading bad actions to be perceived as more intentional than good actions (e.g., Knobe, 2003a, 2010; Pettit & Knobe, 2009). This somewhat controversial claim has generated substantial attention from scholars in many fields, leading to a number of alternative proposals for why the effect occurs (e.g., Adams & Steadman, 2004a, 2004b, 2007; Alicke, 2008; Guglielmo & Malle, 2010; Guglielmo, Monroe, & Malle, 2009; Hindriks, 2014; Lau & Reisenzein, 2016; Laurent, Clark, & Schweitzer, 2015; Machery, 2008; McGuire, 2012; Phelan & Sarkissian, 2008; Scaife & Webber, 2013; Sripada, 2012; Sripada & Konrath, 2011; Uttich & Lombrozo, 2010; Wiland, 2007). Despite this, it has been argued that no single account has yet fully and convincingly explained the effect (Nichols & Ulatowski, 2007). The current proposal—which does not require that moral considerations influence ordinary intuitions about intentionality—builds on past explanations, incorporates existing knowledge about moral reasoning, and provides what we believe is a convincing explanation for the intentionality side-effect effect. In making our case, we (a) challenge a core assumption underlying extant side-effect effect research, (b) identify and examine the critical concepts underlying people's disparate responses to side effect scenarios in helping and harming cases, and (c) experimentally manipulate these concepts to test their effects on intentionality responses. In so doing, we isolate the information that influences people's judgments and construct a new and parsimonious explanation for observed asymmetries in intentionality judgments.

The side effect story outlined earlier (Knobe, 2003a) seems relatively uncomplicated because it is easy to believe that a business leader would value profits over environmental concerns. Yet, it contains a wealth of information about the chairman's foreknowledge, goals, expectations, attitudes, desires, and moral character. Although the only words that differ in the stories regard the valence of side effects, we argue that people interpret the two cases in substantially divergent ways. These differences in interpretation may make it difficult to directly compare participants' responses to intentionality questions across the cases, challenging whether strong inferences can be drawn about people's concepts of intentional action.

For example, it has typically been assumed that seemingly straightforward survey questions about intentionality (e.g., "Did the chairman intentionally help/harm the environment?") are capturing equivalent judgments in both helping and harming cases. However, in our first experiment, we show that outcome valence influences perceivers' understanding of the intentionality questions they are asked. In the harming case, perceivers redefine the question to focus on what the agent knew would happen (i.e., the side effect) when he intentionally acted and whether he deserves to be blamed for acting anyway. In the helping case, foreknowledge is less important because agents deserve no accolades for self-serving actions that also result in incidental benefits they care nothing about (e.g., Bartsch & Young, 2010). Thus, participants confronted with the help case believe they are being asked whether the agent's intentional action was *directed at* helping (i.e., whether he started the program, even partly, in order to help). In short, we show that although both questions seem to be asking the same thing, people in different conditions appear to be answering different questions, with each question focused on a different aspect of intentionality.

Building on this, we then show that when foreknowledge about harm is absent, relatively few people respond that harming was intentional. Yet, even when agents are strongly and actively against harming, if foreknowledge is present, people continue to label harming intentional. On the other hand, although people deny intentionality to helping on the sole basis of foreknowledge, people label helping as intentional at rates similar to the harming case when agents actively want to help, even when helping is not the agents' primary goal.

We are not arguing that the side-effect effect is not real or that moral considerations fail to affect how people respond to intentionality questions in side effect cases. Clearly, these effects exist and tell us important things about how people understand and explain others' morally charged behavior. However, we do question what causes the effect to emerge and whether side effect outcome valence influences lay intuitions about the concept of intentionality. Below, we outline the arguments in support of our case in greater detail.

The Side-Effect Effect and the Moral Influence Hypothesis

Since its initial demonstration (Knobe, 2003a), the side-effect effect has been replicated many times (e.g., Cova, Lantian, & Boudesseul, 2016; Cova & Naar, 2012; Cushman & Mele, 2008; Knobe, 2003b, 2004; Mele & Cushman, 2007; Nichols & Ulatowski, 2007; for reviews, see Feltz, 2007; Knobe, 2010). It has been found across different cultures (e.g., Knobe & Burra, 2006; but see Lau & Reisenzein, 2016) and related to brain function (Ngo et al., 2015). Even children (Leslie, Knobe, & Cohen, 2006; Rakoczy et al., 2015) and professional judges (Kneer & Bourgeois-Gironde, 2017) fall prey to its influence. The effect on intentionality is also not the only side-effect effect; observed effects have involved other mental states such as knowledge (e.g., Beebe & Buckwalter, 2010; Beebe & Jensen, 2012), desire (e.g., Nadelhoffer, 2006; Pettit & Knobe, 2009), judgments of "doing" versus "allowing" (Cushman, Knobe, & Sinnott-Armstrong, 2008), and beliefs about causality (e.g., Knobe & Fraser, 2008). To explain these easily replicable and interesting effects—particularly the effect on how people think about intentional action—a moral influence hypothesis has been proposed. It suggests that moral considerations such as the badness of an outcome pervasively influence lay intuitions about whether a behavior was intentionally performed (e.g., Knobe, 2003a, 2004, 2007, 2010; Knobe & Burra, 2006; Pettit & Knobe, 2009).

In brief, the moral influence hypothesis posits that when evaluating side effect scenarios, people automatically compare agents' perceived attitudes against normative moral defaults (i.e., what one might *expect* agents' attitudes to be or what they *should* be; e.g., Knobe, 2010). The defaults shift as a function of outcome (Figure 1). For a bad outcome, the default is that agents should have at least *some* con-attitude toward it (i.e., they should be somewhat against it), presumably because most people would share this attitude. The default for a good outcome is that agents should have at least *some* pro-attitude

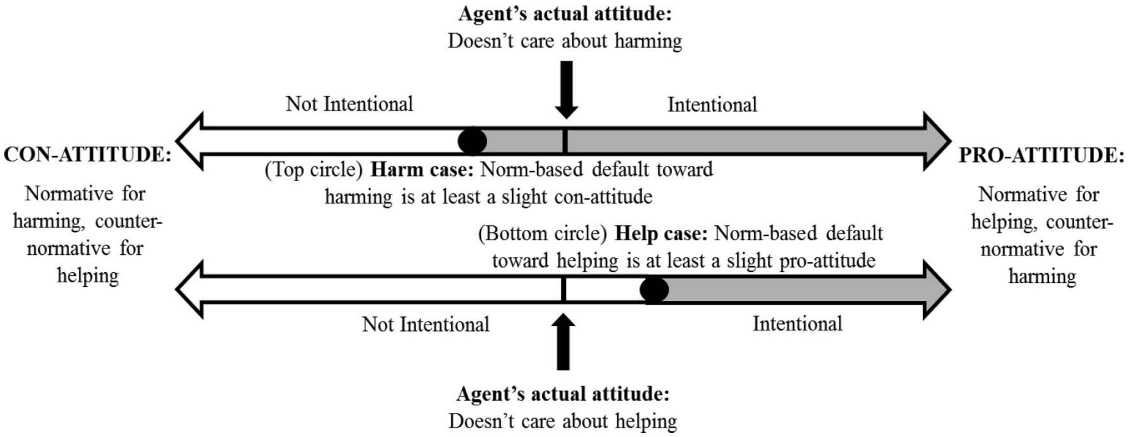
Agent's actual attitude:
Doesn't care about harming
Not Intentional
Intentional
CON-ATTITUDE:
Normative for harming, counter-normative for helping
(Top circle) Harm case: Norm-based default toward harming is at least a slight con-attitude
(Bottom circle) Help case: Norm-based default toward helping is at least a slight pro-attitude
PRO-ATTITUDE:
Normative for helping, counter-normative for harming
Not Intentional
Intentional
Agent's actual attitude:
Doesn't care about helping

Figure 1. Graphic depiction of the moral influence hypothesis (see Knobe, 2010; Pettit & Knobe, 2009).

toward it (i.e., they should be somewhat in favor of it), for the same reason. Judgments about intentionality rely on the location of the agent's perceived attitudes relative to the defaults. When the harming agent's actual attitude is sufficiently "con" (i.e., at or below the bad-outcome default), perceivers should not attribute intentionality to harming; above this default, harming will be seen as intentional. Likewise, when agents' attitudes are sufficiently "pro" (i.e., at or above the good-outcome default), perceivers will attribute intentionality to helping; below this, intentionality will be denied. Because the agents *do not care* about the outcome in either case, the attitudes fall respectively above and below the defaults in the harming and helping cases, leading people to call harming but not helping intentional (Figure 1).

This is an interesting hypothesis, but a few points deserve further consideration. First, it is not clear which attitudes correspond to the normative defaults. Knobe (2010) suggests that defaults represent "a judgment that could be made even in the absence of any information about this specific agent or his behaviors" (p. 328). Thus, defaults might be characterized as decontextualized beliefs about the attitudes most people have toward particular helpful or harmful outcomes, such as those concerning the environment (i.e., widely shared moral norms). Yet, context matters in side-effect effects because intentionality questions are necessarily asked and answered *after* receipt of information about the agent and his behaviors. Are the appropriate defaults therefore people's personal beliefs about *what the agent's attitudes should have been* toward helping or harming or *what type of attitude most people would have* if they were in the same position as the agent? Arguments could be made for either case, but the closest approximation to the moral influence hypothesis is probably the latter judgment. This is because people's personal beliefs are by their nature subjective, involving their own attitudes, as well as complex and affectively-charged observations about the agent's mental states. For example, in the popular chairman scenario, participants' judgments probably take into account the chairman's foreknowledge, goals, attitudes toward profits and the environment, and his decision to act, along with their own feelings about the environment, corporate profit pursuit, and the outcome. On the other hand, considering what most people's attitudes would be if confronted with the same situation should force perceivers into a more objective frame that better matches the defaults proposed by the moral influence hypothesis, which requires no knowledge of the agent and his behaviors. In this case, even if participants anchor on their own attitudes and insufficiently adjust to consider what most people's attitudes might be (e.g., Epley & Gilovich, 2006), estimating these attitudes should result in a closer approximation to what the agent's attitudes reasonably *should* be.

Second, the defaults mean something different across the helping and harming cases, which plausibly contributes to the intentionality effect. Although some companies might include environmental protection as part of their mission, most observers should reasonably assume that profits are a company's main goal (Laufer, 2003).[1] Given this assumption, the harming agent in the chairman scenario faces a trade-off that the helping agent does not (Machery, 2008). Starting the program benefits the company only at a cost to the environment. Not starting it avoids harming, but costs the company. Thus, tension exists between responsibility to the company and obligations to avoid harming the environment, making the default attitude ambiguous.[2] Should the chairman be so against harming the environment that he will sacrifice profits to avoid it? Would most people feel this way?

On the other hand, helping is easy and cost-free, because environmental benefit is only a side effect that does not thwart the primary goal of increasing profits. What possible reason might the chairman have for not starting the program, other than animus toward the environment? In light of this counterfactual, his vocally dismissive attitude about helping likely connotes a true disdain for environmental concerns, resulting in a belief that he is actively against helping and only grudgingly helps in order to increase profits. This proposition is consistent with previous findings regarding perceived desire; in harming cases, desire tends to be rated near scale midpoints (suggesting indifference), but in helping cases it is rated as particularly low (e.g., Ditto, Pizarro, & Tannenbaum,

[1] Even in this case, it would be assumed that unless the company is described as a non-profit organization, one of their main goals would be to make profits.

[2] It would be hard to imagine a company whose main or secondary goal is to cause environmental damage.

2009; Guglielmo & Malle, 2010; Tannenbaum, Ditto, & Pizarro, 2007, as cited in Pettit & Knobe, 2009).

Third, manipulating the chairman's actual attitudes should be reasonably easy and can allow inferences to be made about the causal role of these attitudes on intentionality responses. However, manipulating the defaults—what *most people's* decontextualized attitudes would be—presents a greater challenge, suggesting that a comparison of the agent's actual attitudes to the defaults may matter less than the attitudes alone. If true, when agents have strong con-attitudes toward harming, harming should not be labeled intentional, and when agents have strong pro-attitudes toward helping, people should say they intentionally helped.

One final consideration about the moral influence hypothesis should be noted. Even though the harming agent's action in side effect cases seems obviously blameworthy and the helping agent's action deserves no praise, Knobe (2010) states that, "the present account . . . makes no mention at all of blame" (p. 328; see also p. 324 and Alicke, 2008; Alicke & Rose, 2010). This suggests that neither blame nor praise has a role in how people answer questions about intentionality. We return to this point below in our discussion of why the intentionality side effect is particularly important, both generally and in legal contexts (e.g., Kneer & Bourgeois-Gironde, 2017; Malle & Nelson, 2003), and in our arguments for how foreknowledge, desire, blame, and praise contribute to the emergence of the side-effect effect.

Why Focus on Intentionality?

One reason to study the intentionality side effect is that many implications arise if the previously assumed causal relationship between intentionality and morality is reversible. For example, people's negligent or reckless actions might be viewed as intentional even when no intent to cause harm is present, leading to harsher punishments than would otherwise be the case. It is also the most studied side-effect effect (Knobe, 2010). Furthermore, other mental states implicated in side effects, such as knowledge and desire, are frequently viewed as underpinning intentionality judgments (e.g., Adams, 1986; Heider, 1958; Jones & Davis, 1965; Malle & Knobe, 1997; Shaver, 1985). Most important, intentionality is foundational to social and moral cognition (e.g., Heider, 1958; Malle, 2006; Malle et al., 2001, 2014; Reeder, 2009).

Perceiving that an agent has intentionally caused harm incites moral outrage and a desire for retribution (Darley & Pittman, 2003), making intent and intentionality particularly important inputs to moral judgments and the assignment of blame (e.g., Cushman, 2008; Guglielmo, 2015; Guglielmo et al., 2009; Malle, Guglielmo, & Monroe, 2012, 2014; Shaver, 1985). For example, agents who cause harm intentionally (vs. accidentally) are seen as more responsible and deserving of blame and punishment (Ames & Fiske, 2013; Darley & Pittman, 2003; Malle, 2006; Malle, Guglielmo, & Monroe, 2014; Robinson & Darley, 1995; Schultz & Wright, 1985; Young & Saxe, 2011). Moreover, these concepts figure prominently into how people assign responsibility, blame, and punishment in legal contexts (Malle & Nelson, 2003), evidenced in the differences between manslaughter or reckless manslaughter and first-degree murder (e.g., Model penal code, 1981; N.Y. State Penal Codes 125.15 and 125.27).

Yet, people are still negatively judged for causing unintended (e.g., negligent) harm (e.g., Ames & Fiske, 2013; Nobes, Panagiotaki, & Pawson, 2009; Schultz & Wright, 1985; Schultz, Wright, & Schleifer, 1986). Moreover, people use information about mental states thought to underlie intentionality inferences (i.e., beliefs, desires, awareness) to form judgments about immorality, blame, and punishment when agents cause harm but do not specifically intend to do so (Cushman, 2008; Laurent, Nuñez, & Schweitzer, 2015, 2016; Nuñez, Laurent, & Gray, 2014). This illustrates the importance of these mental states to moral cognition across a variety of cases. As we discuss in the next section, we believe that foreknowledge and desire play particularly important roles in generating side-effect effects. Foreknowledge is important because if an agent does not know that some action will probably lead to an outcome, it makes little sense to say they intentionally acted in order to bring that outcome about (Wiland, 2007). Desire is important for a different reason: It provides a reason (e.g., Malle, 1999) that explains why agents act—because they are trying to bring some outcome about (see also Adams & Steadman, 2004a).

Foreknowledge and Desire in Side-Effect Effects

Because people assign a relatively heavier weight to bad things relative to good (e.g., Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001), rules about not harming are more strictly enforced than rules about helping (e.g., Carnes & Janoff-Bulman, 2012; Janoff-Bulman, Sheikh, & Hepp, 2009). This is probably why people receive substantial blame for the "deliberate indifference" or "gross carelessness" characterizing recklessness or "gross negligence" (e.g., Fitzgerald & Williams, 1962; Stark, 2016). In these cases, agents do not particularly desire or intend harmful consequences but know they are possible and act in a way that risks their occurrence anyway (Garner, 2014). Consistent with this, recklessness has been described as a hybrid state somewhere between negligently caused and intentionally caused harm (Darley & Pittman, 2003); all that is missing is desire and specific intent to harm, leading behaviors to be "treated in many respects as if it [harm] were so intended" (Keeton, Dobbs, Keeton, & Owen, 1984, p. 213). Because of this, people may naturally label harmful, but not helpful side effects as intentional. Furthermore, "recklessly" is used to describe blameworthy actions that lead to socially undesirable consequences, making its use similar to that of "intentionally," a word that most commonly references negative outcomes (Malle, 2006).

This argues for the importance of foreknowledge and its relation to blame in labeling harming as intentional in side effect cases. When asked whether the agent intentionally harmed, the question may be reinterpreted as asking whether the agent deserves blame, because when harm is not intended, blame should principally rely on whether it was foreseeable and preventable (e.g., Malle et al., 2014). That is, denying that the agent intentionally harmed might be interpreted as indicating that he does not deserve to be blamed. Equally or more likely, people may reinterpret the question as asking whether the agent intentionally acted (i.e., in pursuit of a goal), knowing that harm would be a secondary result, rather than as literally asking whether harming was what the agent did intentionally.

We note that others have advanced related arguments. For example, it has been suggested that participants will say an agent intentionally harmed because no other alternatives are provided that would better match their beliefs, such as "knowingly harmed."

Consistent with this, when offered a choice between indicating that an agent intentionally versus knowingly harmed, participants prefer the latter explanation (e.g., Guglielmo & Malle, 2010; Laurent, Clark, & Schweitzer, 2015; see also Adams & Steadman, 2004a, 2004b). However, this does not explain why, when offered a choice between "knowingly, *but not* intentionally harmed" and "knowingly *and* intentionally harmed," a majority of people choose the latter description in morally charged contexts, but not in morally neutral contexts. To explain this finding, Adams and Steadman (2007) argued that even though people can draw distinctions between intentionally and knowingly caused harms, the close linkage between blame and intentionality in natural language makes it difficult to override an inclination to call a blameworthy action intentional.

Our account shares features with these other explanations, but differs in some important ways and can help explain why people prefer "knowingly" to "intentionally," but also prefer "knowingly and intentionally" to "knowingly, but not intentionally." Specifically, we believe that the question about intentionality invites a complex inferential process (e.g., Royzman & Hagan, 2017), prompting people to use available information to try and understand a question asking about the intentionality of an *outcome* rather than the action that causes it (see Laurent, Clark, et al., 2015). This leads them to reintroduce the missing action and interpret the question (in the harming case) as asking whether the agent intentionally acted (i.e., to pursue some goal, such as profits) in full knowledge that this action would lead to a harmful side effect. This interpretation is consistent with participants' self-reported reasons for indicating that the agent in the chairman scenario intentionally harmed: he implemented a program (intentional action), knowing it would harm the environment (foreknown side effect; Nichols & Ulatowski, 2007; e.g., p. 349).

If foreknowledge is enough for ascription of intentionality, why then is side effect helping not labeled intentional? One reason may be that people do not commonly use the word "intentionally" to describe actions with positive, socially-desirable consequences (Malle, 2006). Similarly, there is no equivalent expression to "recklessly" that describes actions leading to unintended but foreseen positive outcomes. What then would it take to label helping intentional? On our view, when side effect outcomes are helpful, people primarily focus on what agents were trying to accomplish when they acted (e.g., Wiland, 2007), and desire information provides a strong clue about an agent's goals. To the extent that intentionality is a concept typically applied to goal-directed behaviors (e.g., Baldwin & Baird, 2001; Baird & Astington, 2004; Behne, Carpenter, Call, & Tomasello, 2005), it makes sense that helping is not labeled intentional. The agent is not *trying* to help the environment and helping does not explain why he acted.

An alternative explanation that also involves both foreknowledge and desire is that there may be more than one concept of intentionality, and that people apply different concepts depending on outcome valence (e.g., Cushman & Mele, 2008; Ditto et al., 2009; Gintis, 2010; Nichols & Ulatowski, 2007; Scanlon, 2010). For bad outcomes, intentionality might be defined as acting with foreknowledge, even when an outcome is not specifically desired or intended. For good outcomes, a definition that includes foreknowledge, desire, intent, and other mental states (e.g., Malle & Knobe, 1997) might be required to attribute intentionality. Although this is possible, we do not believe that multiple definitions of intentionality are required to explain the effect. Instead, we argue that people *reinterpret* intentionality questions differently across cases and answer these questions based on what they think they are being asked.

The Present Research

As discussed previously, the intentionality side-effect effect has been investigated in a variety of contexts, suggesting that research on the topic is of broad interest to a diverse group of people, such as social psychologists and philosophers, or people studying human development, marketing, the brain, or the law. Essentially, because it can inform what we know about how people think about and evaluate the actions of others, it should be of interest to anyone curious about moral judgment or how people think about intentional action.

In the experiments that follow, we first demonstrate how, in the harming case, people redefine intentionality questions as asking about foreknowledge and blame; in the helping case, people interpret the question literally or believe they are being asked about what the agent was trying to accomplish when he acted. We also show that the same concepts are involved in people's judgments of blame and praise. Experiments 2a–2c show that without foreknowledge, few participants respond that agents intentionally harm or help. Finally, Experiments 3a–3c show that when desire to help is increased, the frequency of labeling helping as intentional also increases (see also Guglielmo & Malle, 2010). On the other hand, increasing desire to *avoid harming* does not affect the rates of labeling harming intentional. These last experiments also provide evidence that the labeling of harm as intentional or not does not depend on the relation of agents' attitudes to normative moral defaults.

General Method

All manipulations and measures are disclosed. Verbatim wording of all instructions and measures are provided in an online supplement to this article, which also includes all procedures (available at http://osf.io/4va6j/). Fully de-identified data for all experiments are available at the same location.

Because of the large number of experiments presented, we describe overall sample characteristics here rather than in each method section. Across all experiments, participants were 1529 U.S. residents recruited through Amazon's MTurk (AMT) website and paid a small fee for their participation. This total includes data that are footnoted but not reported in the main text (i.e., pilot tests of Experiments 2c, 3a, and 3c; see footnotes 8 and 10). Sample sizes were: Experiment 1 $n = 121$, Experiment 2 control $n = 53$, Experiment 2c pilot $n = 100$, Experiment 2a $n = 51$, Experiment 2b $n = 51$, Experiment 2c $n = 199$, Experiment 3 control $n = 200$, Experiment 3a pilot $n = 101$, Experiment 3c pilot $n = 48$, Experiment 3a $n = 202$, Experiment 3b $n = 200$, Experiment 3c $n = 203$. All sample sizes were determined in advance by pilot testing for effect sizes, conducting power analyses, or basing sample sizes on past research. In addition, tests of key hypotheses used multiple experiments and replications. No data analyses were performed until target sample sizes were achieved. Prior to data collection, all research was approved by relevant institutional review boards. Informed consent was obtained from all participants prior to participation. All conditions in all experiments were between participants with participants randomly assigned to condition. On the basis of unique AMT identifiers, each participant participated in only one condition of one experiment. At the end of each experiment, participants provided demographic information, received a code for payment, and were thanked.

Sample characteristics were as follows: $M_{\text{age}} = 35.53$, $SD = 11.96$; 51.7% female, 47.9% male (remaining participants reported "other" or "prefer not to disclose"). In Experiments 1, 2a, 2b, and the pilot of Experiment 3c ($N = 324$), a 9-point scale measuring political liberalism/conservatism (1 = *extremely liberal*, 5 = *middle of the road*, 9 = *extremely conservative*) showed that the sample leaned liberal ($M = 4.11$, $SD = 2.17$).[3] Either one or two simple attention check questions were used in each experiment. Few participants answered one or more incorrectly (3.1% or 47/1,529). All analyses were conducted two ways: using all participants and excluding those who failed at least one attention check. No statistical (i.e., significance testing) or conceptual (e.g., direction of frequencies or means) conclusions differed as a function of including/excluding any participants. We therefore retained all participants in reported analyses.

For descriptive purposes, responses in all experiments were recoded to center on scale midpoints. No other data transformations were performed. To conserve space, we frequently use the abbreviation "HH" to refer to "help and/or harm," "helped and/or harmed," "helpful versus harmful," and similar combinations.

Experiment 1

We believe that when making decisions about intentionality of harming, the primary inputs to participants' (yes) responses will involve foreknowledge and counterfactuals about blame (i.e., that denying intentionality to harming indicates that the chairman should not be blamed). In the helping case, inputs to (no) responses should involve the chairman's goals, what he was trying to do, and his reasons for acting. To a lesser extent, they should also be about denial of praise. These differences in focus across conditions should also cause participants to redefine questions about intentionality of HH in different ways. That is, rather than literally interpreting the questions as asking whether the chairman's *intentional action* was to harm or help (see also Laurent, Clark, et al., 2015), they should interpret questions as respectively asking about foreknowledge/blame and reasons/goals. These same inputs and redefinitions should be closely aligned with reasons for assigning blame and denying praise. That is, the harming chairman should be seen as blameworthy because he intentionally acted (i.e., started a program), *knowing* this would lead to harm. The helping chairman should receive little praise because he started the program only as a means to increasing profits, not in order to help the environment.

Method

Information in plain text was presented to participants in both conditions. Words in brackets that are bolded (*bolded and italicized*) were presented only to participants in the helping (harming) condition. Instructions are abbreviated here; complete instructions are provided in the online supplemental material. Unless otherwise noted, responses were on 9-point scales where 1 = "*completely disagree*" and 9 = "*completely agree*."

Participants were presented with either the helping or harming version of the original chairman vignette. Next, they were asked, "Did the chairman intentionally HH the environment?" Response options were "no" or "yes." Because our interest was solely in normative responses (i.e., "no" in the helping condition and "yes" in the harming condition), only those participants who responded normatively (the vast majority of participants) were prompted: "Using the provided scales, indicate your agreement with the following statements."

The first set of questions focused on participants' explanations of their intentionality responses. Participants rated reasons that completed the prompt: "I responded that the chairman [**did not intentionally help**] [***intentionally harmed***] the environment because . . ."

Trying: "when he acted he [**was not**] [***was***] trying to HH the environment.

Goal: "when he acted, his goal was [**not**] to try and HH the environment."

Reason: "[**helping the environment was not the reason the chairman acted; he acted to increase profits**] [***harming the environment was the reason the chairman acted***]."

Meant: "the chairman **[didn't mean]** [***meant***] to HH the environment."

Wanted: "the chairman **[didn't want]** [***wanted***] to HH the environment."

Knew: "**[the chairman's intentional action was not to help the environment; it was to start a program he knew would help it]** [***although the chairman's intentional action was not to harm the environment, he started a program he knew would harm it***]."

Credit/Blame: "saying he **[intentionally helped]** [***did not intentionally harm***] the environment would be like saying he [**deserves credit**] [***does not deserve blame***] for HH it."

Participants then rated statements about the meaning of the intentionality question by responding to the prompt: "What did you interpret the question about whether the chairman intentionally HH the environment to mean? The question was asking . . ."

Trying: "whether the chairman, when he acted, was trying to HH the environment."

Goal: "whether the chairman's goal, when he acted, was to try and HH the environment."

Reason: "whether the chairman's reason for acting was to HH the environment."

Wanted: "whether the chairman wanted to HH the environment."

Intentional: "whether the chairman's intentional action was to HH the environment."

Knew: "whether the chairman knew, when he acted, that the environment would be HH."

Credit/Blame: "whether the chairman deserved **[credit]** [***blame***] for HH the environment."

Participants were then asked, "If you had to select one of the following only, which option best captures the meaning of the question you responded to about whether the chairman intention-

[3] Across all experiments, no significant differences in age or ideology were found as a function of experimental condition, $ts(1{,}527 \text{ and } 322) = 0.04$ and 1.13, $ps = .972$ and .258. Similarly, no association between self-reported gender and condition was found, $\chi(3, N = 1{,}529) = 2.39$, $p = .495$.

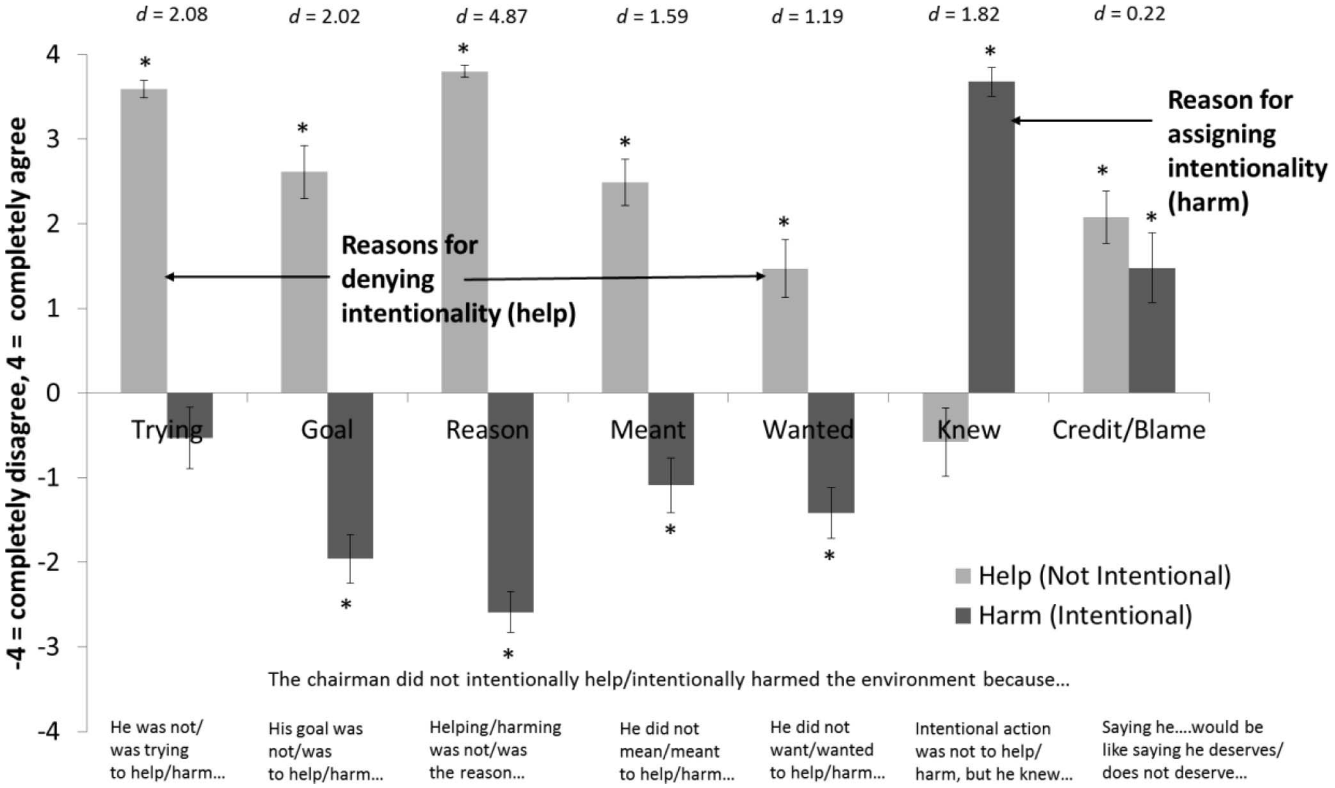

Figure 2. Condition-based means regarding perceived reasons for denying/assigning intentionality to helping/harming in Experiment 1. Except for "credit/blame" ($p = .239$), all condition-based differences were significant, $p < .001$. Error bars represent $\pm SE_{\text{Mean}}$. Effect size *d* for condition-based differences is given above comparisons. * indicates a mean that significantly differs from its scale midpoint, $p \leq .001$.

ally HH the environment? The question was asking . . ." Response options were the same as above and participants could select only one option.

Participants in both conditions (including non-normative responders to the intentionality question) were then asked, "Do you think the chairman deserves **credit** [***blame***] for HH the environment?" Response options were "no" or "yes." As above, only people who provided normative responses (i.e., "no" for credit or "yes" for blame) were asked, "On the previous page, you indicated that you **[do not]** think the chairman deserves **[credit]** [***blame***] for HH the environment. Why did you respond in this way?" They then rated reasons that completed the prompt: "I think the chairman **[does not deserve credit]** [***deserves blame***] because . . ."

Trying: "he was [**not**] trying to HH the environment."

Goal: "it was [**not**] his goal to HH the environment."

Reason: "**[helping the environment was not the reason he acted; he just wanted to increase profits]** [***harming the environment was the reason he acted***]."

Wanted: "he [**did not want**] [***wanted***] to HH the environment."

Intentional: "his intentional action was [**not**] literally to HH the environment."

Know-Intent: "it does not matter whether **[he knew the environment would be helped when he acted; what matters is whether it was his intention to help it]** [***it was his intention to harm the environment; what matters is that he knew it would be harmed when he acted***]."

Participants were then asked, "If you had to select one of the following reasons for why the chairman **[does not deserve credit]** [***deserves blame***] for HH the environment, which would it be? He **[does not deserve credit]** [***deserves blame***] because . . ." Response options were the same as above and participants could select only one option.

Results and Discussion

Intentionality responses. As is typically found, few participants responded that the chairman intentionally helped (2/61). Almost all responded that he intentionally harmed (56/60), $\chi^2 = 98.29$, $p < .001$, $\varphi = 0.90$.

Perceived inputs to intentionality responses. Including only those who respectively responded "no" and "yes" in help and harm conditions, significant condition-based mean differences in reasons for choosing a particular response emerged for all variables except "credit/blame" ($p = .239$), $ts(113) = 6.33$ ("wanted") to 26.32 ("reason"), $ps < .001$ (Figure 2).[4]

[4] Effect sizes for all condition-based mean differences are reported in accompanying figures. For variables where only data summaries (i.e., means and *SD* or *SE*) are reported in text, we also provide (in the online supplemental material) graphical summaries of the frequency of participants who endorsed each response as a function of condition.

Assuming that responses significantly above or below midpoints indicate disagreement or agreement (vs. ambivalence or uncertainty), we also tested whether, within conditions, means significantly differed from scale midpoints. In the helping condition, participants clearly believed that trying, goals, reasons for acting, and what the agent meant to do and wanted to do (all significantly above scale midpoints) were important inputs to their decisions to deny intentionality. In the harm condition, with the exception of "trying" (which did not differ from its scale midpoint) people disagreed that these same inputs were important for their decisions about intentionality of harming. Instead, the overwhelming reason they endorsed was that even though the chairman's intentional action was not to harm the environment, he acted (i.e., started a program) while *knowing* it would harm the environment ($d = 2.97$, relative to scale midpoint). Although participants agreed that the potential denial of blame in the harm condition and deservingness of credit in the help condition influenced their responses, responses to this question did not differ across condition, showing that assignment of blame or praise was viewed as a relatively important reason for responses in both cases.

Perceived meaning of intentionality question. Means for all definitions of the intentionality question significantly differed across conditions, $ts(113) = 3.97$ ("wanted") to 8.06 ("knew"), ps $< .001$ (Figure 3). In the help condition, participants' responses suggested they either reinterpreted the question about intentionality as asking about what the agent was trying to accomplish (e.g., "was he trying to help," or "was helping why he acted?") or took the question at face value as asking whether the chairman's intentional action was to help. In the harm condition, there was ambivalence about these definitions. Instead, participants redefined the question as asking about the chairman's foreknowledge when he acted ($d = 1.82$) and his deservingness of blame for harming ($d = 1.16$).

In the help condition, over 88% (combined) of participants selected "trying" (14/59), "goal" (16/59), "reason" (12/59), or a literal interpretation ("intentional," 10/59) as the best definition of the intentionality question. In the harm condition, over 83% (combined) selected "know" (34/56) or "blame" (13/56). Only 3/56 participants believed that the question should be interpreted literally, $\chi^2(6) = 68.80$, $p < .001$, Cramer's V $= 0.77$.

This suggests that across conditions, participants were answering what they thought were different questions (Figure 3), even though the wording of the intentionality question was exactly the same except for "harm" and "help." Participants in the help condition mostly thought the question was asking about the agent's goals, his reasons for acting, or what he was trying to do when he acted. Some also thought the question was literally asking whether helping was the chairman's intentional action. Participants in the harm condition overwhelmingly thought they were being asked about what the chairman knew when he intentionally acted, but

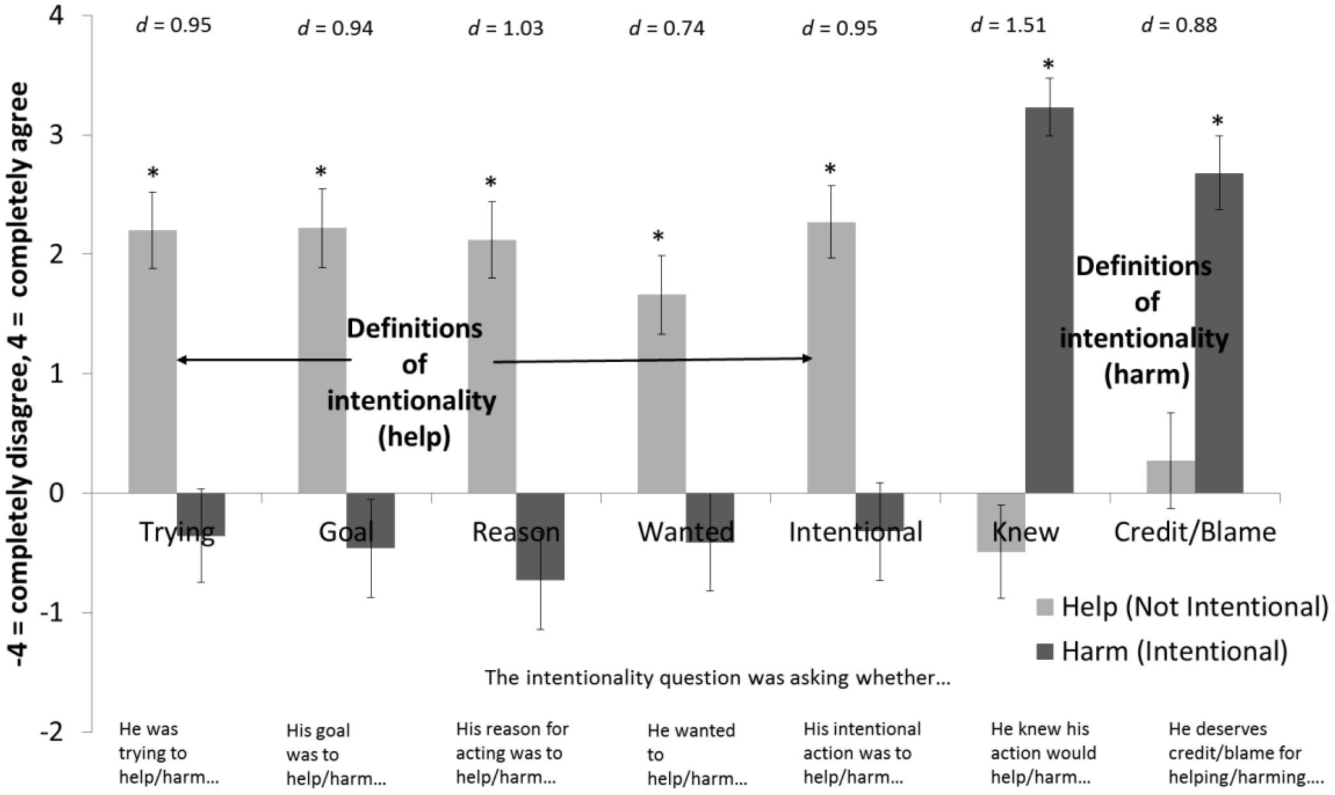

Figure 3. Condition-based means regarding definitions of the intentionality question in Experiment 1. All condition-based mean differences were significant, $p < .001$. Error bars represent ± SE_{Mean}. Effect size d for condition-based differences is given above comparisons. * indicates a mean that significantly differs from its scale midpoint, $p < .001$.

also whether he deserved blame for acting, given this foreknowledge. Participants in this condition did not agree that the question was literally asking whether his intentional action was to harm.

Credit/blame responses. Few participants (13/61) gave the chairman credit for the helpful outcome. Most (59/60) assigned blame for harming, $\chi^2(1) = 74.47$, $p < .001$, $\varphi = 0.78$.

Perceived inputs to credit/blame responses. Including only participants who respectively responded "no" (does not deserve credit) and "yes" (deserves blame), ratings for all questions significantly differed across conditions, *t*s(105) = 4.89 ("know-intent") to 19.17 ("reason"), *p*s < .001 (Figure 4). When forced to choose one response as the best reason for denying credit in the help condition, the most frequently chosen (31/52) response was that helping the environment was not the reason the chairman acted—he was simply pursuing a goal of profits. Other important reasons were that he was not trying to help (7/52) and that his intentions and not his foresight mattered (6/52). In the harm condition, the overwhelmingly best reason (53/55) was that his intentions did not matter, his foresight did, $\chi^2(5) = 80.09$, $p < .001$, Cramer's V = 0.87.

From an examination of the means (Figure 4) and participants' choices of the best reasons for denying credit, it is clear that these decisions used the same perceived predicates as denying intentionality: The agent was not trying to help, and helping was not the reason for which he acted. That is, the agent's foreknowledge was of little importance, but the goal that motivated his action mattered a great deal. In the harm condition, agreement was low that these factors were important for blame. Instead, blame rested almost solely on the importance of foresight relative to intent, which mirrors the findings regarding why participants responded that he intentionally harmed and how they defined the question about intentionality of harming.

Overall, participants do not appear to believe that harming was the chairman's intentional action (which is what the question, on its face, appears to be asking); instead, their reason for indicating that the chairman intentionally harmed was that he intentionally acted, knowing harm would result. For side effect helping, participants thought foreknowledge was not nearly as important as what goal the chairman was pursuing, which explains why he acted. Thus, removing foreknowledge should greatly decrease the labeling of harm as intentional, and increasing perceptions that the chairman wants to help—even if helping is not his primary goal—should increase the labeling of helping as intentional.

Experiments 2a–2c examine whether removing foreknowledge decreases the labeling of harming as intentional. Following this, Experiments 3a–3c examine whether increasing desire increases the labeling of helping as intentional. Each of these experiments also tests whether asymmetries in labeling helping and harming as intentional persist even under these conditions.

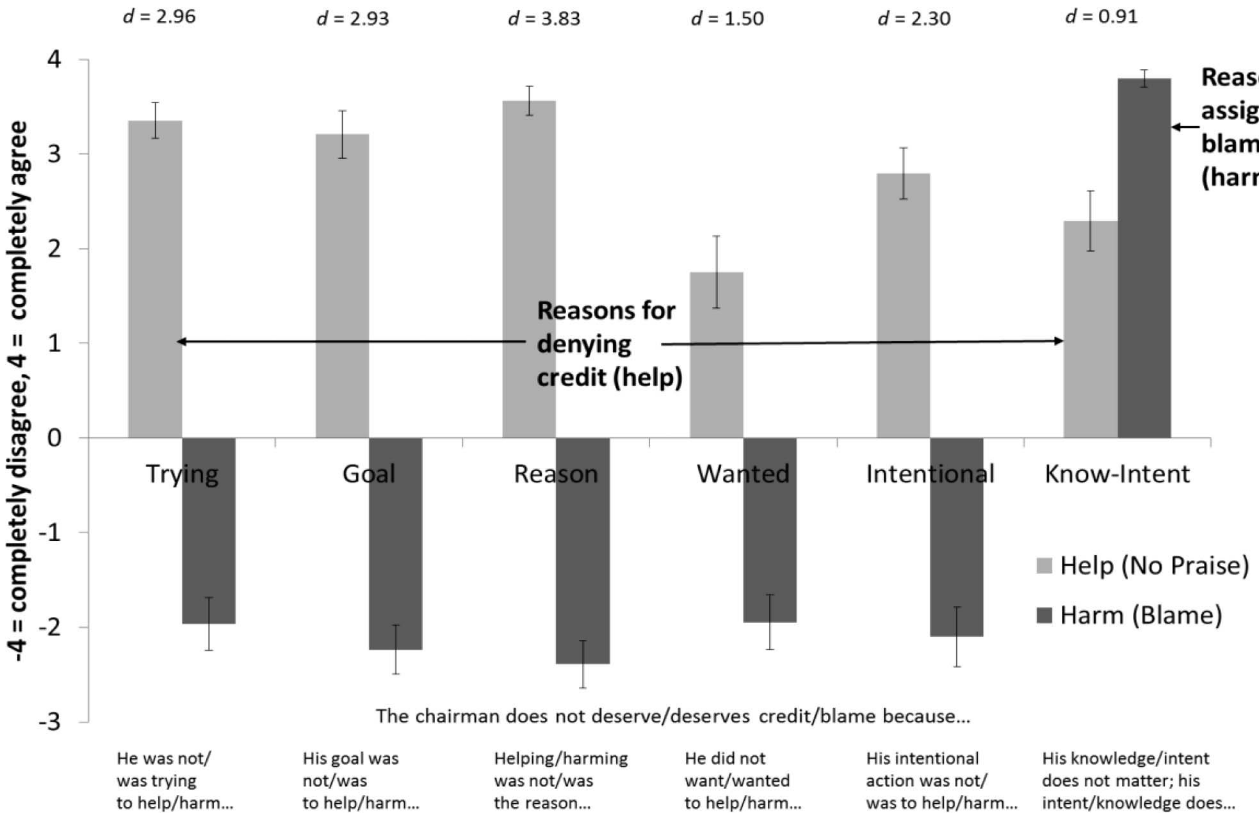

Figure 4. Condition-based means regarding perceived reasons to not praise or to blame in Experiment 1. All condition-based mean differences were significant, $p < .001$. Error bars represent ± SE_{Mean}. Effect size *d* for condition-based differences is given above comparisons. All means significantly differed from scale midpoints, $p < .001$.

Experiments 2a–2c

Experiments 2a–2c examine whether the asymmetry in intentionality responses emerges when foreknowledge about environmental side effects is absent. We expected that relative to control (i.e., the original chairman vignette), removing foreknowledge would decrease the labeling of harm as intentional. We also hypothesized that when foreknowledge was absent, no asymmetries in labeling helping and harming as intentional would emerge.

In Experiment 2c, we also measured perceptions of the agents' pro-con attitudes toward HH, participants' personal beliefs about what the agents' attitudes toward HH *should* be, and their general beliefs about what sorts of attitudes *most people* would have about HH. Previously, we proposed that general beliefs are probably closest to the defaults outlined by Knobe (2010), because even if biased by participants' own attitudes, they should represent participants' best estimates about generic (i.e., decontextualized) moral norms involving helping and harming the environment. On the other hand, participants' personal beliefs are likely to be highly subjective. Because of this, we expected that personal defaults would be more than "a little bit" or "slightly" (Knobe, 2010, p. 327) pro for helping or con for harming. That is, because helping is so easy, personal defaults in the help case should be strong pro-attitudes. Similarly, because on average, people should be quite opposed to harming, personal defaults should be strong con-attitudes. However, it is difficult to know what precise values would convincingly represent demarcation points between "very strong" and "somewhat strong" or "slightly strong" pro or con attitudes. Thus, we hypothesized that personal defaults would be stronger than general defaults.

The moral influence hypothesis predicts that if the harming chairman's attitude is less con-harming (i.e., more proharming) than the defaults, people will label harming intentional (see Figure 1). Our competing hypothesis was based on the importance of foreknowledge for evaluating harmful side effects. We predicted that in the harm condition, the chairman's attitudes would be clearly less con than defaults, but that nevertheless, relatively few people would label harming intentional.

Method

In a control experiment, people were presented with the original chairman vignette. This was used to examine whether removing foreknowledge in Experiments 2a–2b impacted rates of labeling helping and/or harming as intentional.[5] In Experiment 2a, the chairman is described as caring only about profits, not whether the environment is helped or harmed (i.e., as in the original chairman vignette). Then, unexpectedly, the environment is respectively harmed and helped. In Experiment 2b, the chairman's attitudes about the environment are not described and he is not told about any consequence to the environment. Thus, he cannot possibly know it will be helped or harmed when he decides whether to start the program. Because in Experiment 2a, the chairman's attitude does not directly correspond with the effect on the environment (e.g., he does not care about *helping* and then the environment is *harmed*), Experiment 2c addresses this potential issue. The chairman is again described as not caring whether the environment is helped or harmed. Because he doesn't care, he is not told about any consequence to the environment, and again cannot possibly know that the environment will be helped or harmed when he starts the program. The full text of all vignettes is provided in the online supplemental material.

In all experiments, after being presented with the story, participants were asked whether the chairman intentionally HH the environment.[6] In Experiment 3c, we asked (as a manipulation check), "When he decided to start the program, did the chairman of the board know the environment would be HH?" Response options for both questions were "no" or "yes." Also in Experiment 3c, we asked, "If the chairman had known that the environment would be HH, would he have wanted to HH it?" (1 = *not at all*, 9 = *absolutely*) and, "What do you think the chairman's attitude would have been toward HH the environment, had he known it would be HH? He would have been . . ." (1 = *strongly opposed to it*, 9 = *strongly in favor of it*). These items were aggregated to form a composite pro-con attitude measure ($r = .80$). Two questions asked about participants' personal/general beliefs about what the chairman's/most people's attitudes should have/would have been: "What do you think the chairman's attitude *should have been* toward HH the environment? He should have been . . .;" and "In a similar situation, what type of attitude would *most people have* toward HH the environment? Most people would be . . ." (1 = *strongly opposed to it*, 9 = *strongly in favor of it*).

Results and Discussion

In the control experiment, no participants indicated that helping was intentional (0/27). Most (23/26) responded that the chairman intentionally harmed, $\chi^2 = 41.20$, $p < .001$, $\varphi = .89$. In Experiment 2a, almost no participants indicated that the chairman intentionally helped (1/25) or harmed (4/26), $\chi^2 = 1.87$, $p = .172$, Fisher's exact test, $p = .350$, $\varphi = .17$. Similarly, in Experiment 2b, almost no participants responded that the chairman intentionally helped (2/25) or harmed (2/26), $\chi^2 = 0.002$, $p = .967$, Fisher's exact test, $p = 1.0$, $\varphi = .006$. Relative to control, removing foreknowledge did not impact the frequency of labeling helping intentional in Experiments 2a and 2b, respectively, $\chi^2 = 1.10$ and 2.25, *p*s = .294 and .134, Fisher's exact tests = .481 and .226, $\varphi =$ .15 and .21.[7] As hypothesized, however, removing foreknowledge did reduce the labeling of harming as intentional in the same experiments (i.e., 2a and 2b), respectively, $\chi^2 = 27.81$ and 33.97, *p*s < .001, $\varphi = .73$ and .81.

In Experiment 2c, the manipulation check confirmed that few people thought the chairman knew the environment would be helped (18/94) or harmed (21/105), $\chi^2 = 0.23$, $p = .880$, $\varphi = .01$.

[5] Because all conditions of all experiments were between-participants and comparisons were of frequencies of labeling HH intentional, we used the same control conditions as comparators for Experiments 2a and 2b. Because the sample size of our control was smaller than that of Experiment 2c, we did not perform any comparisons in this case (although, for example, we could have compared frequencies against those of Experiment 1).

[6] In Experiment 2 control, Experiments 2a and 2b, and the pilot test of Experiment 2c, participants in help/harm conditions were respectively asked questions about morality/immorality, praise/blame, and greed. Responses to these questions, although interesting, are not central to our primary hypotheses. Thus, these analyses are presented only in the online supplemental material.

[7] Fisher's exact tests are also reported for Experiments 2a, 2b, and help condition comparisons because of low expected frequencies (< 5) in some cells.

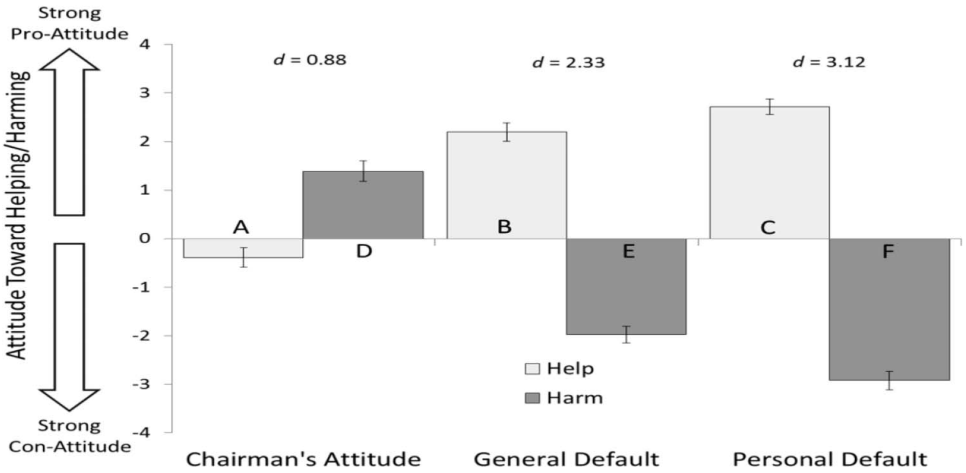

Figure 5. Comparisons of perceived attitudes with general and personal defaults in Experiment 2c. Within conditions, means marked with different letters are significantly different using paired-samples *t*-tests, $p < .05$. The letters A, B, and C indicate help condition comparisons and the letters D, E, and F indicate harm condition comparisons. All between-conditions means are significantly different, *p*s < .001. Between-condition effect sizes are given above comparisons. Bars are ± SE_{Mean}.

Replicating the same effects on intentionality as Experiments 2a and 2b, few participants thought the chairman intentionally helped (13/94) or harmed (15/105), $\chi^2 = 0.01$, $p = .926$, $\varphi = .01$.[8] As expected, perceived attitudes, general defaults, and personal defaults were more pro in the help than in the harm condition, ts(197) > 6.16, *p*s < .001 (Figure 5). Next, we used repeated measures *t* tests within conditions (respectively in help and harm conditions, $df = 93$ and 104) to compare the defaults with one another and to examine perceived attitudes relative to the defaults.

As hypothesized, personal defaults were significantly stronger (i.e., more prohelping and con-harming) than general defaults (particularly in the harm condition), respectively in help and harm conditions, ts = 2.24 and 4.53, *p*s = .028 and < .001. Theoretically, the more important finding was that in the harm condition, perceptions of the chairman's attitudes were far above both defaults (i.e., were not at or below the con-attitude defaults; in fact, people assumed he would be somewhat in favor *of* harming), ts ≥ 13.08, *p*s < .001. In the help condition, the chairman's attitude was significantly below defaults, as might be expected, ts > 9.97, *p*s < .001.

Results of Experiments 2a–2c were consistent with our hypotheses. Foreknowledge appears to be particularly important in harming side effect cases. Removing foreknowledge causes most people to stop labeling harming as intentional, which results in similar numbers of people labeling harming and helping as intentional (and no significant asymmetries). Because participants already fail to label helping intentional in the standard side effect case, removing foreknowledge does not change how they respond.

Another interesting finding was that the helping chairman's attitudes were rated as somewhat ambivalent, but the harming chairman's attitude seemed to be actively in favor of harming. This differs from the usual case where desire to help is seen as particularly low and desire to harm is seen as ambivalent. Future research might investigate why this was the case here. Of note, this finding suggests in another way that in the harm case, the chairman's pro- or con-attitudes (whether on their own, or relative to defaults) do not drive how people respond about intentionality. That is, people did not say he intentionally harmed even though he appeared to be somewhat in favor of doing so.

More important, a prediction from the moral influence model (e.g., Knobe, 2010) was at odds with what was found in the harm condition. Despite perceiving the chairman's attitudes as far less con than defaults in the harming case—which, according to this model, should have led participants to believe the chairman intentionally harmed—very few people labeled harming as intentional. Finally, because outcome valence differed across conditions in each experiment but the frequency of intentionality responses did not, it is clear that outcome valence alone does not drive the intentionality side-effect effect. The answer appears to be more nuanced.

Experiments 3a–3c

To increase perceptions of desire (to help and not harm), in each experiment, the chairman was described as strongly proenvironment (i.e., strongly in favor of helping and strongly against harming). Our primary hypothesis was that increasing desire to help would primarily increase the frequency of labeling helping intentional (see also Guglielmo & Malle, 2010). Because of the presence of foreknowledge, we expected that majorities of participants would still label harming intentional even though the agent was clearly opposed to harming.

We took several different approaches to testing this hypothesis. Experiment 3a described a company in serious financial trouble. Thus, starting a profit-increasing, environment-helping program represents an easy choice in the help condition, because the chairman can pursue his primary goal (increasing profits) but also bring

[8] In a pilot test of Experiment 2c, few people said the chairman knew the environment would be helped (9/50) or harmed (14/50), or that the chairman intentionally helped (11/50) or harmed (15/50), respectively, $\chi^2 =$ 1.41 and 0.83, $p = .235$ and .362, $\varphi = .12$ and .09.

about a desirable side effect. In the harm condition, the trade-off represents a difficult choice, because the chairman is against harming the environment but genuinely needs to start the program. In Experiment 3b, no mention of financial distress was mentioned. Again, profits were described as the primary consideration, in order to strengthen perception that helping was an additional goal in the helping condition (i.e., profits motivated starting the program, but were not essential to the company's future). In the harming condition, no trade-off was necessary; since profits were not essential, the chairman *could have* decided to not start the program. In Experiment 3c, the company was again described as in financial distress. To create a trade-off in the help condition and also increase perception that the chairman's main goal was to help, the decision to help required sacrifice: the chairman was told that starting the program would not guarantee profits and was more likely to decrease profits. In the harm condition, the chairman was again faced with the choice of starting a program or perhaps going bankrupt.

Similar to Experiment 2c, each experiment measured perceptions of the agents' pro-con attitudes toward helping and harming, personal defaults regarding what his attitudes should be, and general defaults about what most people's attitudes would be in the same situation. We again considered general defaults to be the most appropriate comparator and hypothesized that personal defaults would be consistently stronger than general defaults. We also expected that because personal defaults might be particularly strong (as in Experiment 2c), it would be very unlikely for perceptions of the chairman's attitudes—no matter how clearly pro-helping or con-harming—to ever surpass these values. However, in the harm conditions, we expected the chairman's con-attitudes to be similar to or more con than the general defaults. Despite this, we expected people to continue to label harming intentional because the chairman did intentionally act, knowing that environmental harm would be a secondary result.

To help confirm that in Experiments 3a and 3b, people believed that increasing profits is what motivated the chairman's action (i.e., to assure that helping remained a side effect), we measured perceptions regarding why the agents had acted—to increase profits versus to HH the environment. In Experiment 3c, we expected people to believe that the helping chairman's goal was to help, because helping required a potential sacrifice. We also measured the extent to which people thought the agents were *trying* to HH the environment. We predicted that in all harm conditions, people would not think the chairman was actively trying to harm the environment. In the help condition of Experiment 3c, we expected people to strongly believe the chairman was trying to help the environment. In the other help conditions, given the chairman's proenvironment attitudes, we expected them to believe, at least in part, that he was trying to help.

Method

In a control experiment, people were presented with the original chairman vignette. This was included to test whether increasing pro-attitudes toward helping would increase the frequency of labeling helping intentional, but that increasing con-attitudes toward harming would not decrease the frequency of labeling harming intentional. In Experiments 3a-3c, participants read one of three different chairman vignettes. The chairman in each case was described as strongly proenvironment (pro-helping or against harming). In Experiments 3a and 3c, the chairman's company was described as in serious financial trouble and in danger of bankruptcy, with profits badly needed. In Experiment 3b, no mention of the company's financial situation was made. In Experiments 3a and 3b, helping and harming were described as (fortunate and unfortunate) side effects. In the help condition of Experiment 3c, increased profits were described as only a slight possibility and loss of profits as more likely. Because the chairman was willing to risk losing money and started the program anyway, helping was probably viewed as the primary reason he started the program.

Similar to Experiment 2c, participants were asked the following questions, in order[9]: "Did the chairman intentionally HH the environment?" (no or yes). "Did the chairman want to HH the environment?" (1 = *not at all*, 9 = *absolutely*) and, "Rate what you think the chairman's attitude was toward HH the environment. He was . . ." (1 = *strongly opposed to it*, 9 = *strongly in favor of it*). These items were aggregated to form composite pro-con attitude measures (respectively, in control and Experiments 3a–3c, *r*s = .84, .87, .81, and .88). Personal and general beliefs about what the chairman's/most people's attitudes should/would be were captured using single items: "What do you think the chairman's attitude *should have been* toward HH the environment? He should have been . . ." and, "In a similar situation, what type of attitude would *most people have* toward HH the environment? Most people would be . . ." (1 = *strongly opposed to it*, 9 = *strongly in favor of it*). Participants were then asked two questions as manipulation checks: "What was the chairman's main goal in starting the program? His main goal was . . ." and "The chairman started the program in order . . ." (1 = *to HH the environment*, 9 = *to increase profits*). These were aggregated into measures of goal pursuit (respectively in control and Experiments 3a–3c, *r* = .87, .82, .75, and .91). Finally, one question asked, "By starting the program, was the chairman *trying* to HH the environment?" (1 = *absolutely not*, 9 = *absolutely*).

Results

Goals and trying (manipulation checks). Results are summarized in Table 1. Relative to the chairman in all helping conditions (except control), the chairman in all harming conditions seemed to be motivated more by profits. Similarly, except in Experiment 3 control, the helping chairman was always rated as trying to help to a greater extent than the harming chairman was rated as trying to harm. In the help conditions of Experiments 3a and 3b and in all harming conditions, participants believed the chairman's primary goal was increasing profits, as indicated by means significantly above scale midpoints, $p\text{s} < .001$. In Experiment 3c, the helping chairman appeared to be motivated more by helping the environment than by profits, $p < .001$. In the helping condition of Experiment 3a, the chairman did not seem to be trying to help in particular, $p = .163$. In Experiments 3b and 3c, the helping chairman seemed to be trying to help, $p\text{s} < .001$. The harming chairman was never perceived as trying to harm, $p\text{s} < .001$.

[9] As in earlier experiments, questions were asked about morality/immorality, praise/blame, and greed. Analyses of these measures are reported in the online supplemental material.

Table 1
Experiment 3 Control and Experiments 3a–3c: Descriptive Statistics for Help and Harm Conditions, Tests Comparing Means (on Goals and Trying) and Frequencies (of Intentionality Responses) Across Help and Harm Conditions, and Tests Comparing Frequencies of Intentionality Responses (Within Help and Harm Conditions) Against Control in Experiments 3a–3c

	Help (n = 100)		Harm (n = 100)				
(Control) Original vignette	*M*	*SD*	*M*	*SD*	*t*(198)	*p*	*d*
Goal: HH vs. Profits	3.55	1.19	3.29	1.30	1.45	.149	.21
Trying to HH	−2.85	2.24	−.61	2.43	6.77	<.001	.96
	No	Yes	No	Yes	χ^2	*p*	φ
Intentional?	84	16	12	88	103.85	<.001	.72
	Help (n = 104)		Harm (n = 98)				
Experiment 3a	*M*	*SD*	*M*	*SD*	*t*(200)	*p*	*d*
Goal: HH vs. Profits	2.67	1.65	3.49	1.29	3.92	<.001	.55
Trying to HH	**.33**	2.37	−2.07	2.20	7.44	<.001	1.05
	No	Yes	No	Yes	χ^2	*p*	φ
Intentional?	43	61	18	80	12.64	<.001	.25
Experiment 3a Help Condition vs. Help Control					39.47	<.001	.44
Experiment 3a Harm Condition vs. Harm Control					1.56	.212	.09
	Help (n = 100)		Harm (n = 100)				
Experiment 3b	*M*	*SD*	*M*	*SD*	*t*(198)	*p*	*d*
Goal: HH vs. Profits	1.39	2.12	3.44	1.16	8.47	<.001	1.09
Trying to HH	1.22	2.23	−1.68	2.36	8.93	<.001	1.26
	No	Yes	No	Yes	χ^2	*p*	φ
Intentional?	22	78	16	84	1.17	.279	.08
Experiment 3b Help Condition vs. Help Control					77.16	<.001	.62
Experiment 3b Harm Condition vs. Harm Control					.66	.415	.06
	Help (n = 100)		Harm (n = 103)				
Experiment 3c	*M*	*SD*	*M*	*SD*	*t*(201)	*p*	*d*
Goal: HH vs. Profits	−1.28	2.42	3.50	1.22	17.89	<.001	2.49
Trying to HH	2.44	1.95	−1.65	2.44	13.17	<.001	1.85
	No	Yes	No	Yes	χ^2	*p*	φ
Intentional?	9	91	13	90	.69	.407	.06
Experiment 3c Help Condition vs. Help Control					113.05	<.001	.75
Experiment 3c Harm Condition vs. Harm Control					.02	.893	.01

Note. HH = help and/or harm," "helped and/or harmed," "helpful versus harmful," and similar combinations. Except where bolded (p = .163), means significantly differ from scale midpoints, *p*s < .05. All variables (except intentionality) were measured on 9-point scales and centered on scale midpoints. In Experiment 3a, companies were in financial distress and profits were framed as the main goal. In Experiment 3b, no mention of the company's financial situation was made and profits were again the main goal. In Experiment 3c, companies were in financial distress and the chairman in the helping condition had to risk losing profits to help.

Intentionality responses. Table 1 provides frequencies of labeling helping and harming as intentional in all experiments. Except in the control experiment, majorities labeled both helping and harming as intentional. Significant asymmetries (in the traditional direction) only emerged in Experiment 3a and control.[10] In the other two experiments, the asymmetries were not significant. As hypothesized, manipulating attitudes increased labeling of helping as intentional relative to control in all experiments (all *p*s < .001); these manipulations had no impact on labeling of harming as intentional, *p*s > .211.

Chairman attitudes and moral defaults. As depicted in Figure 6 (which provides all between-condition effect sizes), significant condition-based differences were found in all experiments regarding the chairman's attitudes toward HH, personal beliefs about what his attitudes should have been, and general attitudes about what types of attitude most people would have toward HH in the same situation, all *t*s > 4.59, all *p*s < .001. Demonstrating the helping chairman's proattitudes toward helping and the harming chairman's con-attitudes toward harming Experiments 3a–3c, all respective attitude means were significantly above and below scale midpoints, *t*s > 6.75, *p*s < .001. This was reversed in the control experiment using the original chairman vignette, *t*s > 8.58, *p*s < .001. In the help conditions, all personal and general moral defaults were above scale midpoints, *t*s > 10.42, *p*s < .001. One exception

[10] Consistent with the finding for Experiment 3a, in the pilot of this experiment, about half of participants labeled helping intentional (24/52) and a slight majority labeled harming intentional (33/49), χ^2 = 4.61, p = .032, φ = .21. In the pilot test of Experiment 3c, almost all participants labeled helping intentional (20/23) and a smaller majority labeled harming intentional (15/25). The asymmetry was significantly reversed, χ^2 = 4.41, p = .036, φ = .30.

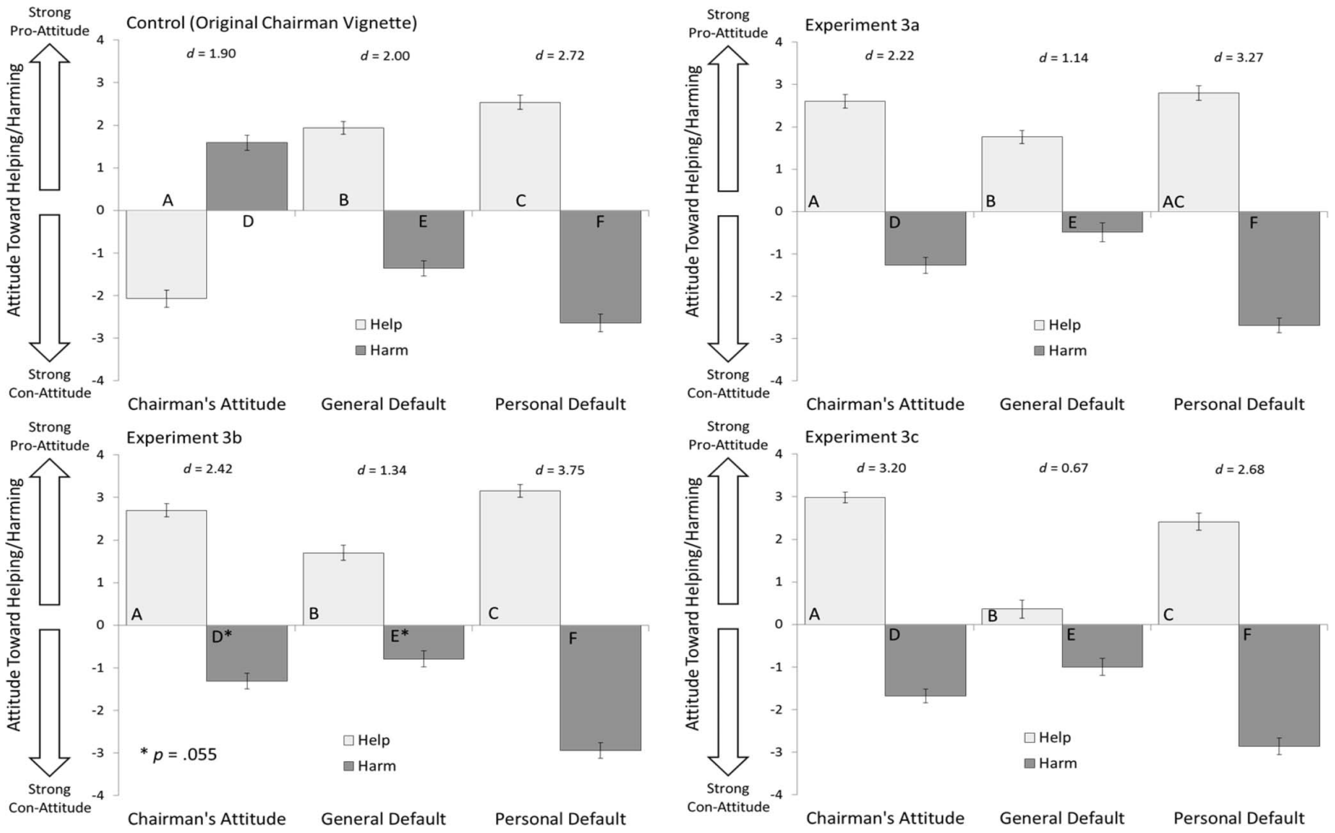

Figure 6. Comparisons of perceived attitudes with general and personal defaults in Experiment 3 control (top left), and Experiments 3a (top right), 3b (bottom left), and 3c (bottom right). Within conditions, means marked with different letters are significantly different using paired-samples *t* tests, $p < .01$. The letters A, B, and C indicate help condition comparisons and the letters D, E, and F indicate harm condition comparisons. All between-conditions means are significantly different, *p*s < .001. Between-condition effect sizes are given above comparisons. Bars are ± SE_{Mean}.

is that in Experiment 3c—when helping was potentially costly—the general default did not significantly differ from the scale midpoint, $t = 1.75$, $p = .083$. In the harm conditions, the personal and general moral defaults were always significantly below scale midpoints, indicating that the chairman should not and that most people in his situation would not want to harm the environment, *t*s > 2.21, *p*s < .030. Paired-samples *t* tests also showed that, as hypothesized, personal defaults were consistently higher than general defaults in both help (*t*s > 3.28, *p*s ≤ .001) and harm (*t*s > 6.01, *p*s < .001) conditions, suggesting that personal defaults represented particularly strong pro- and con-attitudes.

A final set of tests examined whether the chairman's perceived pro- or con-attitudes were significantly different from personal or general defaults. To examine this question, we used paired-samples *t* tests to compare perceived attitudes with personal and general defaults within each condition of each experiment. With the exception of the control experiment, in all help conditions, the chairman's attitudes were significantly more pro than the general default, *p*s < .001. In Experiment 3a, they were not significantly different from the personal default ($p = .271$). In Experiments 3b and 3c, respectively, they were significantly less pro ($p < .001$) and more pro ($p = .006$) than personal defaults. In the harm conditions, the chairman's attitudes were significantly more con than general defaults in Experiments 3a ($p = .006$) and Experiment 3c ($p = .002$); in Experiment 3b, this difference was marginally significant, $p = .055$. The harming chairman's attitudes were always significantly less con than personal defaults, *p*s < .001 (see Figure 6).

Discussion

As proposed, relative to control, manipulating the chairman's attitudes to be in favor of helping caused an increase in the frequency of labeling helping as intentional. Instead of most people denying intentionality to helping as they typically do (and did, in the helping condition of the control experiment), majorities indicated that the chairman intentionally helped. Rates of labeling helping as intentional were high enough that no intentionality asymmetries were observed in Experiments 3b (when helping remained a side effect) and 3c (when helping appeared to be the chairman's primary goal).

On the other hand, con-attitudes toward harming had little, if any, influence on labeling harm as intentional. Rates of labeling harming intentional remained quite high across all experiments and did not significantly differ from control in any experiment, even though the agents' attitudes were consistently perceived as against harming. This is somewhat puzzling when considering results from a conceptually similar experiment reported in Guglielmo and Malle (2010, Study 4a). These authors found that when a CEO expressed regret for having to harm the environment, rates of labeling harming intentional were substantially reduced (to around 40%). One speculative possibility for the difference regards subtle differences in how desire was operationalized in the two cases. In Guglielmo and Malle (2010), the CEO's simple expression of regret may have led some participants to assume the intentionality question was asking about what motivated his action, similar to participants in the help condition of our Experiment 1. In the current experiments, the agents never directly expressed regret over having to cause harm (although regretful attitudes were likely inferred). Moreover, in Experiments 3a and 3c, the harming chairman (explicitly described as pro-environment) faced a clear tradeoff between trying to avoid bankruptcy and protecting the environment. It is also worth noting that con-attitudes were never perceived as particularly strong in spite of the chairman's avowed attitudes. Likely, this is because in each case, despite the chairman's pro-environment stance, he decided to implement the program. Thus, even when he might have been pragmatically justified in making this decision, perceptions of his attitudes took his choice into account, making it probable that the only way participants would have believed the chairman was completely against harming would have been if he decided to not start the program. Thus, the way desire was operationalized in the current studies may have encouraged participants to consider that after deliberating, the chairman decided to intentionally act, even knowing it would lead to harm. If so, they may have assumed the question was asking just that: whether, despite his attitudes, he intentionally acted, knowing it would lead to harm (e.g., Experiment 1). Future research might examine this possibility directly.

A clear pattern also emerged regarding defaults. As we argued above, the general defaults were much closer to defaults proposed by Knobe (2010; Pettit & Knobe, 2009). In the harm cases in particular, general defaults appeared to represent "slight" con-attitudes toward harming, even in the control experiment. Compared with these, the personal defaults were always stronger (and appeared quite strong). In fact, personal defaults were so strongly against harming that the only way the chairman reasonably could have surpassed them is if he decided to not start the program. Even in this hypothetical case, for perceived con-attitudes to be descriptively lower than the observed average for personal defaults would have required almost complete agreement from participants that his attitudes were as strongly con-harming as was possible using the given scale (i.e., below 2 on the 9-point scales). On the other hand, even though personal defaults were stronger than general defaults in the help cases, general defaults for helping were strong overall, at least when helping was not associated with any costs (Experiments 3a and 3b). However, when helping was costly (Experiment 3c), the general default was less extreme and did not differ from its scale midpoint.

In many ways, the general defaults seem to signify reasonable assessments of what the agents' attitudes should be. That is, most people would probably agree that the agents should be against harming and in favor of helping, but would also agree that their actual attitudes should take into account the situations they find themselves in. If caring for the environment is an important goal, but increasing profits is the chairman's main goal, then being strongly in favor of starting a program that will lead to profits and will also help the environment is sensible. Realistically however, if starting a program that will help the environment will also increase the probability that the chairman's company will go bankrupt, his zeal for helping should be substantially tempered. Similarly, he should generally be against starting a program that will harm the environment. Yet, if the only way his company will survive is by staring that program, it is understandable if his attitude about harming is not as negative as it would ordinarily be.

If our argument about the general defaults being an appropriate comparator is correct, then it is difficult for the moral influence hypothesis to explain why people continued to label harming intentional (or why only a slight majority labeled helping intentional in Experiment 3a). That is, the chairman's con-attitude was consistently more con than the general default in each experiment, which should have led most people to indicate that the chairman did not intentionally harm. And if one proposes that the personal defaults are more appropriate—even though, short of not starting the program, the harming chairman's attitudes would be unlikely to ever reach these levels—then the frequency at which people labeled helping intentional in Experiment 3b becomes difficult to explain. Moreover, it is difficult to explain using this hypothesis why most participants did *not* label harming intentional in Experiment 2c, even though the chairman's attitudes were far less con-harming than both defaults.

We think these findings are easier to explain within the framework we have proposed—that desire has less impact on calling harming intentional because when the chairman's actions lead to side effect harm, people are less concerned with his attitudes than with how he acted, given his foreknowledge. And in helping cases, people care less about foreknowledge; what they mostly care about is the chairman's goals and what he was trying to accomplish when he acted.

General Discussion

At the outset, we argued that understanding the intentionality side-effect effect is particularly important, largely because inferences regarding intentionality are foundational to social cognition (e.g., Malle et al., 2001), moral evaluations (e.g., Ames & Fiske, 2013; Malle et al., 2014; Reeder, 2009), and the law (e.g., Darley & Pittman, 2003; Malle, 2006; Malle & Nelson, 2003). Like others, we agree that moral considerations play a substantial role in how people respond to questions about intentionality in side-effect cases, but we disagree that it does so by influencing how people apply the concept of intentionality to action (e.g., Knobe, 2010). Instead, we have proposed that because of differences in how people think about actions that lead to harm and benefit, foreknowledge information is useful for evaluating actions leading to harmful side effects, and reasons for acting—such as what the agent was trying to accomplish—are important for evaluating actions leading to helpful side effects. Because different mental states are important for understanding and appraising agents' actions in different side effect cases, people assume questions about intentional harming and helping (i.e., the side-effect outcomes) are asking about the role these distinct mental states played in informing actions, and answer accordingly. A series of experiments supported this position.

Our hypothesis led to a prediction that people's intentionality responses in harm versus help conditions would respectively be predicated on whether an agent acted in full knowledge of a harmful outcome (i.e., foreknowledge focus) versus whether he acted in order to bring about a helpful outcome (i.e., goal focus). Accordingly, we expected that participants would redefine questions about intentional harming and helping to focus on these concepts, and that the same inputs would be respectively important for their blame and praise decisions. Experiment 1 confirmed that this was the case. In response to a question about whether the chairman intentionally harmed or helped, people believed they were being asked quite different questions (see also Adams & Steadman, 2004a, 2004b; Laurent, Clark, et al., 2015). In the harm condition, the question, "Did the chairman intentionally harm the environment?" was not interpreted literally, and participants did not agree that the question was asking whether harming was the chairman's intentional action. Instead, participants thought the question meant, "Did the chairman know, when he acted, that the environment would be harmed?" or, "Did the chairman deserve blame for harming the environment?" In the help condition, although some took the question at face value—and responded that the chairman's intentional action was not to help—most redefined it in terms of the chairman's goals and reasons for acting. In short, in the harm condition, it was not the agent's goals or intentions that mattered for intentionality and blame responses; it was the agent's intentional action, undertaken with foreknowledge of harm. In the help condition, what mattered was the agent's reason for acting, not what he knew might be an additional consequence of it.

Because of the importance of foreknowledge in the harming case, we predicted that when agents did not know their actions would lead to harm, people would not label harming intentional. We did not expect the absence of foreknowledge to impact the already low rates at which helping is labeled intentional. Our next set of experiments confirmed these predictions. Compared with the same conditions in a control experiment, removing foreknowledge caused fewer people to respond that chairmen intentionally harmed, but did not influence rates of labeling helping as intentional, and rates of labeling the outcome intentional were similar across both conditions. This finding held when agents' attitudes toward the environment were not mentioned (Experiment 2b) and when agents cared only about profits and not at all about the environment (Experiments 2a and 2c). It also remained true when one outcome was foreseen and another occurred (Experiment 2a) and when foreknowledge was not possible because agents weren't told about environmental impacts (Experiments 2b and 2c). In addition, Experiment 2c demonstrated that, at least in this case, participants do not label harming intentional even when the agent's attitudes about harming are far below general or personal default values, which is the outcome predicted by the moral influence model. Overall, responses were consistent with our explanation. If participants in help conditions thought the intentionality question was asking whether the chairman acted in order to help, the correct response would be no, because he did not even know the environment would be helped. If participants in harm conditions thought the question was asking whether he intentionally acted, knowing the environment would be harmed, the correct answer would again be no, because he did not know it would.

Finally, we proposed that if reasons for acting are important determinants to labeling helping as intentional, then if agents seem to genuinely want to help—even if helping is not their main goal and remains a side effect—helping should be more frequently labeled as intentional. On the other hand, desire information should matter less in harming cases because it is primarily agents' foreknowledge and not their attitudes toward harming that causes people to label harming intentional. Three experiments supported both of these ideas. Relative to uncaring attitude conditions in a control experiment using the original chairman vignettes, increasing pro-attitudes toward the environment increased the frequency of labeling helping intentional in each experiment, but did not significantly impact the frequency of labeling harming intentional. Moreover, harming continued to be labeled as intentional even when the agents' attitudes were slightly (Experiment 3b) or significantly below (Experiments 3a and 3c) general defaults. Again, this pattern of responses is consistent with our explanation. In help conditions, participants tended to respond that the chairman intentionally helped because they perceived helping to be, at least in part, the goal of his action (i.e., to some extent, he was trying to help, even if his actions primarily served another goal). In harm conditions, the chairman intentionally harmed because he intentionally acted—even if he might have preferred not to—knowing that harm would result. Finally, combined with the findings from Experiments 2a–2c, these experiments demonstrated that it is not outcome valence per se that drives side-effect effects, because in all but one of these experiments, determinations about intentionality did not vary across outcome type.

Conclusions

As we noted at the outset and reiterate here, we are not arguing that side-effect effects are not real, or that moral considerations play no role in people's decision-making about side-effect cases. Because people overwhelmingly (and one could argue, rightly) believe that it is wrong and blameworthy to knowingly harm and not particularly right or praiseworthy to uncaringly help, people's moral intuitions are of paramount importance in how they evaluate side-effect cases. Despite this, we are less certain that people's responses to questions regarding intentionality actually reflect divergent intuitions about what constitutes an intentional action or require the application of different meanings of intentionality in different cases. Instead, we believe that moral considerations affect what mental states (i.e., associated with the agent's intentional action) people *believe are important to evaluate* across the two cases, despite being asked about intentionality: what he knew in the harm case and why he acted in the help case.

As a corollary of this natural shift in focus, when agents knowingly harm, people will label harming intentional, demonstrating the importance of the agents' foreknowledge to their responses. This may be particularly likely when agents seem unconcerned about harming, but should represent the majority response even when agents appear to be actively *against* harming. On the other hand, foreknowledge on its own should matter much less in the helping case—particularly since foreknown and uncaring helping is already denied intentionality. In this case, the agent's attitudes are of paramount importance. When helping seems to represent, even in part, the *reason* for an agent's action (i.e., demonstrating desire and intent), responses that they have intentionally helped should increase. This should be true even when helping is a side effect and not the agent's primary goal.

Further research might be able to provide additional evidence for our arguments. For example, jointly manipulating levels of foreknowledge and desire and crossing these variables could more directly test additional predictions, such as whether foreknowledge only impacts the labeling of helping as intentional when desire is also present. It might also be useful to test statistical mediation of outcomes on intentionality responses through perceptions of agents' attitudes, moral judgments of the agent or his actions, determinations regarding the praiseworthiness and blameworthiness of their actions, and perceptions of foreknowledge and reasons for acting. Given the current set of findings, we would predict that foreknowledge, reasons for acting, and praise/blame would mediate intentionality judgments, but that attitudes (i.e., desire to help or harm) and moral judgments would not.

Limitations and Future Directions

A few potential limitations are worth noting, as are several open questions that future research might address. One limitation is that all of the current studies were based around Knobe's (2003a) chairman vignette. The decision to test our hypotheses using this example alone was motivated by several reasons. First, this brief but fascinating story was the first used to document the side-effect effect, and may be the most widely cited example of it in the literature. Second, the chairman vignette consistently exerts strong condition-based effects on people's responses to questions about intentionality, which in theory would make it harder rather than easier to show how these effects can be negated. Finally and most importantly, because we were trying to carefully control a number of potential confounds (e.g., the chairman's attitudes in Experiments 2a–2c; attitudes, trade-offs, and whether the secondary outcome remained a side effect in Experiments 3a–3c), we reasoned that modifications should be made to the same vignette, to allow direct comparisons between studies and to not introduce a new variable (e.g., vignette) that could represent an alternative explanation for our findings. Despite this decision, we believe that using a broader range of vignettes will help push research in this area forward in important ways, such as by providing insight into the conditions under which strong side-effect effects are more or less likely to emerge (e.g., Lau & Reisenzein, 2016; Nadelhoffer, 2006).

Another limitation is that we relied exclusively on between-participants designs. This decision was also made for a few reasons. First, in the real world, on encountering an example of an agent who acts while knowing their action might lead to harm (or more rarely, benefit), it seems unlikely that a person would spontaneously compare this action with a counterfactual case that leads to the opposite effect but holds all other factors constant. Second, with a few exceptions (e.g., Nichols & Ulatowski, 2007), most research on side-effect effects has also relied on between-participants designs. Our intention was to use a methodology similar to those that are typically used to demonstrate the effect, in order to capture people's isolated judgments about each type of case, without inviting consideration of other possibilities. However, future research might test some of the ideas we outline using repeated-measures designs.

In addition to these limitations, other open questions are worth considering, such as whether the framework we propose can be easily extended to other types of side-effect effects (e.g., on foreknowledge, causality, etc.). For some types of side effects, such as causality, we think it can and we have begun to collect data to test this hypothesis. So far, the data appear consistent with the idea that people believe questions about causes are asking about more than just literal causes, particularly because causal language is used in different ways depending on the type of outcome (e.g., causing harm vs. causing help). For other effects, such as on foreknowledge, other types of explanations (e.g., involving base rates and real-world exemplars) might help describe why these effects are found. Again, our ongoing work is testing these ideas and we are finding that people think about foreknown harm and benefit quite differently. As we have proposed here for the intentionality side effect, we believe that moral reasoning is intimately connected to differences in how people think about these other concepts. Yet, we also believe that moral considerations probably do not affect people's core understanding of the concepts.

Ultimately, it may be impossible to prove whether our explanation for the side-effect effect is the correct one, or whether people's actual intuitions about intentional action and other concepts are affected by moral considerations such as outcome valence or the relation of observed attitudes to default moral attitudes. Reasons for the ambiguity include the difficulty of manipulating default attitudes and of uncovering people's potentially complex intuitions without relying on their verbal reports. Despite this, we believe the evidence we have offered, which is based on a solid theoretical foundation, argues for our interpretation of the side-effect effect.

Context

The first author conceived of these ideas while thinking about the mental states typically assumed to underlie how people reason about intentional action (e.g., knowledge, desire, intentions), and how each of these influence moral decision-making in cases where agents do not specifically intend harm. Considering the first author's prior work on similar issues (e.g., regarding perception of negligence and other work on the side-effect effect), the idea that the same mental states might be relevant to the side-effect effect arose naturally. Overall, the research program of the first (and second) author has been increasingly focused on differences in how people reason about blameworthy and praiseworthy actions, and how these two forms of judgment may be applied in different ways, for different reasons.

References

Adams, F. (1986). Intention and intentional action: The simple view. *Mind & Language, 1,* 281–301. http://dx.doi.org/10.1111/j.1468-0017.1986.tb00327.x

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

Adams, F., & Steadman, A. (2004b). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis, 64,* 173–181. http://dx.doi.org/10.1093/analys/64.2.173

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). Aldershot, United Kingdom: Ashgate Publishing.

Alicke, M. D. (2008). Blaming badly. *Journal of Cognition and Culture, 8,* 179–186. http://dx.doi.org/10.1163/156770908X289279

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

Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science, 24,* 1755–1762. http://dx.doi.org/10.1177/0956797613480507

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development, 2004,* 37–49. http://dx.doi.org/10.1002/cd.96

Baldwin, D. A., & Baird, J. A. (2001). Discerning intentions in dynamic human action. *Trends in Cognitive Sciences, 5,* 171–178. http://dx.doi.org/10.1016/S1364-6613(00)01615-6

Bartsch, K., & Young, T. (2010). Reasoning asymmetries do not invalidate theory-theory. *Behavioral and Brain Sciences, 33,* 331–332. http://dx.doi.org/10.1017/S0140525X10001688

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323–370. http://dx.doi.org/10.1037/1089-2680.5.4.323

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language, 25,* 474–498. http://dx.doi.org/10.1111/j.1468-0017.2010.01398.x

Beebe, J. R., & Jensen, M. (2012). Surprising connections between knowledge and action: The robustness of the epistemic side-effect effect. *Philosophical Psychology, 25,* 689–715. http://dx.doi.org/10.1080/09515089.2011.622439

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology, 41,* 328–337. http://dx.doi.org/10.1037/0012-1649.41.2.328

Carnes, N., & Janoff-Bulman, R. (2012). Harm, help, and the nature of (im) moral (in) action. *Psychological Inquiry, 23,* 137–142. http://dx.doi.org/10.1080/1047840X.2012.667768

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,* 1295–1308. http://dx.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,* 837–854. http://dx.doi.org/10.1080/09515089.2011.622363

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108,* 353–380. http://dx.doi.org/10.1016/j.cognition.2008.03.006

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108,* 281–289. http://dx.doi.org/10.1016/j.cognition.2008.02.005

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

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

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. *Psychology of Learning and Motivation, 50,* 307–338. http://dx.doi.org/10.1016/S0079-7421(08)00410-6

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17,* 311–318. http://dx.doi.org/10.1111/j.1467-9280.2006.01704.x

Feltz, A. (2007). The Knobe effect: A brief overview. *Journal of Mind and Behavior, 28,* 265–277.

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

Garner, B. A. (2014). *Black's law dictionary* (10th ed.). St. Paul, MN: West Publishing Company.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56,* 165–193. http://dx.doi.org/10.1016/0010-0277(95)00661-H

Gintis, H. (2010). Modalities of word usage in intentionality and causality. *Behavioral and Brain Sciences, 33,* 336–337. http://dx.doi.org/10.1017/S0140525X10001731

Guglielmo, S. (2015). Moral judgment as information processing: An integrative review. *Frontiers in Psychology, 6,* 1637. http://dx.doi.org/10.3389/fpsyg.2015.01637

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,* 1635–1647. http://dx.doi.org/10.1177/0146167210386733

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 52,* 449–466. http://dx.doi.org/10.1080/00201740903302600

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

Heider, F. (1958). *The psychology of interpersonal relations*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1037/10628-000

Hindriks, F. (2014). Normativity in action: How to explain the Knobe effect and its relatives. *Mind & Language, 29,* 51–72. http://dx.doi.org/10.1111/mila.12041

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

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* (pp. 219–266). Orlando, FL: Academic Press. http://dx.doi.org/10.1016/S0065-2601(08)60107-0

Keeton, W., Dobbs, D., Keeton, R., & Owen, D. (1984). *Prosser and Keeton on the Law of Torts* (5th ed.). St. Paul, MN: West Publishing Company.

Kneer, M., & Bourgeois-Gironde, S. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition, 169,* 139–146. http://dx.doi.org/10.1016/j.cognition.2017.08.008

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis, 63,* 190–194. http://dx.doi.org/10.1093/analys/63.3.190

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16,* 309–324. http://dx.doi.org/10.1080/09515080307771

Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis, 64,* 181–187. http://dx.doi.org/10.1093/analys/64.2.181

Knobe, J. (2007). Acting intentionally and acting for a reason. *Journal of Theoretical and Philosophical Psychology, 27,* 119–122. http://dx.doi.org/10.1037/h0091286

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

Knobe, J., & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture, 6,* 113–132. http://dx.doi.org/10.1163/156853706776931222

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (pp. 441–447). Cambridge, MA: MIT Press.

Lau, S., & Reisenzein, R. (2016). Evidence for the context dependence of the side-effect effect. *Journal of Cognition and Culture, 16,* 267–293. http://dx.doi.org/10.1163/15685373-12342180

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

Laurent, S. M., Clark, B. A., & 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,* 18–36. http://dx.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. http://dx.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,* 1271–1288. http://dx.doi.org/10.1080/02699931.2015.1058242

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect. *Psychological Science, 17,* 421–427. http://dx.doi.org/10.1111/j.1467-9280.2006.01722.x

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

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3,* 23–48. http://dx.doi.org/10.1207/s15327957pspr0301_2

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

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2012). Moral, cognitive, and social: The nature of blame. In J. Forgas, K. Fiedler, & C. Sedikides (Eds.), *Social thinking and interpersonal behavior* (pp. 313–332). New York, NY: Taylor & Francis.

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

Malle, B. F., & Hodges, S. D. (Eds.). (2005). *Other minds: How humans bridge the divide between self and others*. New York, NY: Guilford Press.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33,* 101–121. http://dx.doi.org/10.1006/jesp.1996.1314

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

Malle, B. F., & Nelson, S. E. (2003). Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences & the Law, 21,* 563–580. http://dx.doi.org/10.1002/bsl.554

McGuire, J. M. (2012). Side-effect actions, acting for a reason, and acting intentionally. *Philosophical Explorations, 15,* 317–333. http://dx.doi.org/10.1080/13869795.2012.696130

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

Model Penal Code. (1981). Article 210.

Murder in the First Degree, N. Y. Penal Codes § 125.15 and § 125.27.

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

Ngo, L., Kelly, M., Coutlee, C. G., Carter, R. M., Sinnott-Armstrong, W., & Huettel, S. A. (2015). Two distinct moral mechanisms for ascribing and denying intentionality. *Scientific Reports, 5,* 17390. http://dx.doi.org/10.1038/srep17390

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

Nobes, G., Panagiotaki, G., & Pawson, C. (2009). The influence of negligence, intention, and outcome on children's moral judgments. *Journal of Experimental Child Psychology, 104,* 382–397. http://dx.doi.org/10.1016/j.jecp.2009.08.001

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,* 55–65. http://dx.doi.org/10.1002/acp.2957

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

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. http://dx.doi.org/10.1007/s11098-006-9047-y

Rakoczy, H., Behne, T., Clüver, A., Dallmann, S., Weidner, S., & Waldmann, M. R. (2015). The side-effect effect in children is robust and not specific to the moral status of action effects. *PLoS ONE, 10,* e0132933. http://dx.doi.org/10.1371/journal.pone.0132933

Reeder, G. D. (2009). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological Inquiry, 20,* 1–18. http://dx.doi.org/10.1080/10478400802615744

Robinson, P. H., & Darley, J. M. (1995). *Justice, liability, and blame: Community views and the criminal law*. Boulder, CO: Westview Press.

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

Rosset, E., & Rottman, J. (2014). The big "whoops!' in the study of intentional behavior: An appeal for a new framework in understanding human actions. *Journal of Cognition and Culture, 14,* 27–39. http://dx.doi.org/10.1163/15685373-12342108

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). New York, NY: Routledge.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5,* 296–320. http://dx.doi.org/10.1207/S15327957PSPR0504_2

Scaife, R., & Webber, J. (2013). Intentional side-effects of action. *Journal of Moral Philosophy, 10,* 179–203. http://dx.doi.org/10.1163/17455243-4681004

Scanlon, T. M. (2010). Ambiguity of "intention." *Behavioral and Brain Sciences, 33,* 348–349. http://dx.doi.org/10.1017/S0140525X10001858

Schultz, T. R., & Wright, K. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science, 17,* 97–108. http://dx.doi.org/10.1037/h0080138

Schultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development, 57,* 177–184. http://dx.doi.org/10.2307/1130649

Shaver, K. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4612-5094-4

Sripada, C. S. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology, 48,* 232–238. http://dx.doi.org/10.1016/j.jesp.2011.07.008

Sripada, C. S., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind & Language, 26,* 353–380. http://dx.doi.org/10.1111/j.1468-0017.2011.01421.x

Stark, F. (2016). *Culpable carelessness: Recklessness and negligence in the criminal law*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139855945

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

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition, 116,* 87–100. http://dx.doi.org/10.1016/j.cognition.2010.04.003

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

Wiland, E. (2007). Intentional action and “in order to.” *Journal of Theoretical and Philosophical Psychology, 27,* 113–118. http://dx.doi.org/10.1037/h0091285

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120,* 202–214. http://dx.doi.org/10.1016/j.cognition.2011.04.005

Received July 6, 2018
Revision received November 12, 2018
Accepted November 27, 2018 ■