Can the Knobe Effect Be Explained Away? Methodological Controversies in the Study of the Relationship Between Intentionality and Morality

Florian Cova¹, Anthony Lantian², and Jordane Boudesseul³

Abstract
Based on the “Knobe Effect,” Knobe has argued that moral evaluations can influence intentionality judgments. However, two methodological objections have been raised against this claim: first, that participants’ answers do not accurately reflect what they think and, second, that the Knobe Effect can be fully explained by non-moral factors, such as the agent’s desires or beliefs. In this article, we discuss these two methodological objections to the existence of the Knobe Effect and provide new evidence that moral evaluations can shape intentionality judgments. First, Study 1 shows that standard measures of intentionality do not overestimate participants’ intentionality judgments. Second, Studies 2 and 3 suggest that participants’ moral evaluations still mediate the impact of positive versus negative side-effects on judgments about intentional action, even when taking into account a whole range of non-moral factors. Results suggest that moral evaluations play an irreducible role in shaping our judgments about intentional action.

Keywords
ethics/morality, morality, social cognition, social judgment

Received August 16, 2014; revision accepted May 28, 2016

How people evaluate the moral responsibility of agents and attribute blame and praise to them on this basis has been thoroughly investigated in social psychology. It is now widely accepted that an agent’s responsibility primarily depends on the specific content of her mental states, and particularly on whether the agent intentionally produced a given outcome (Buon, Jacob, Loissel, & Dupoux, 2013; Cushman, 2008; Young, Cushman, Hauser, & Saxe, 2007). Consequently, prominent theories of blame and moral responsibility in moral and social psychology have taken the agent’s intentions as one of the key components of moral responsibility (Cushman, 2008; Heider, 1958; Shaver, 1985; Weiner, 1995). Moreover, focusing on the agent’s intentions is widely considered as a mark of a fully developed adult moral understanding (Cushman, Shekoff, Wharton, & Carey, 2013; Piaget, 1932/2000).

Thus, it seems uncontroversial that information about whether an agent intentionally brought about a given outcome is relevant to a certain kind of moral evaluation, that is, attributions of moral responsibility (to put it schematically, intentionality → moral evaluation). However, in the past decade, a growing body of research has suggested that the reverse might also be true and that certain kinds of moral evaluations might actually shape whether people interpret an agent’s behavior as intentional or not (to put it schematically, moral evaluations → intentionality).

The Knobe Effect
The suggestion that moral evaluations might shape our understanding of actions as intentional or not might appear surprising at first, given that the concept of “intentional action” is traditionally viewed as a descriptive concept, describing a certain kind of relationship between an agent’s mental states and his behavior. For example, Malle and Knobe (1997) argued that for people to consider that an agent intentionally produced a given outcome, five components had to be present: (a) the agent’s desire for this outcome, (b) the belief that his action would produce this outcome, (c) the intention to perform the required action, (d) awareness of the

¹University of Geneva, Switzerland
²University of Reims Champagne-Ardenne, France
³University Grenoble Alpes, Saint-Martin-d’Heres, France

Corresponding Author: Florian Cova, Swiss Center for Affective Sciences, University of Geneva, 7 rue des Battoirs, 1205 Geneva, Switzerland. Email: florian.cova@gmail.com
action being performed, and (e) a certain degree of skill required to perform the action. All these conditions can be considered objective facts about the agent and her mental states and do not require any moral evaluation from the part of the observer to be appreciated.

However, Knobe (2003) obtained results suggesting that, contrary to what was widely assumed, moral evaluations do play a role in our judgments about whether an action counts as intentional. Consider the following vignette:

*Harm* case—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*.

In this case, Knobe found that 82% of the people surveyed answered “yes” to the question, “Did the chairman of the board intentionally harm the environment?” When given the same vignette, but this time with all the occurrences of the word “harm” changed into “help” (the *Help* case), only 23% of participants answered positively when asked if the chairman of the board intentionally helped the environment.

What is puzzling in this asymmetry is that the two cases do not seem to differ from each other in any of the five dimensions deemed relevant by Malle and Knobe (1997): In both cases, the chairman did not intend to either harm or help the environment, nor did he desire to bring such a side-effect. This suggests that there are other factors driving intentionality judgments. One intriguing explanation is that the asymmetry in intentionality judgments is instead driven by a difference in the moral properties of these two cases. Some have argued that what drove the asymmetry was a difference in the agent’s *moral responsibility* (Nadelhoffer, 2006; Wright & Bengson, 2009): While the chairman deserves *blame* for harming the environment in the first case, he does not deserve *praise* for helping the environment in the second case. Others have suggested that what drives the asymmetry in intentionality judgments is the outcome’s *moral valence*: While harming the environment is a *bad* outcome, helping the environment is a *good* outcome (Knobe & Mendlov, 2004).

The claim that moral evaluations can shape intentionality judgments has been thought to have far-reaching implications. On the practical side, some have interpreted these results as showing that people are biased by their moral commitments and concluded that we should worry about jurors’ decisions and impartiality and take the appropriate measures to correct for these biases (Adams, 2015; Nadelhoffer, 2006). Others, on the contrary, have concluded that Knobe’s results show that there is a gap between law and the way we ordinarily think about responsibility and have argued that we should change the law accordingly (Duff, 2015; Kobick, 2010).

On the theoretical side, Knobe has used these results as a starting point to propose a radically original and different conception of folk psychology. Folk psychology has traditionally been considered as having a purely descriptive function and as making use of the same methods one might find in a formal scientific investigation. However, Knobe considers that his results suggest an alternative picture, according to which moral considerations actually figure in the fundamental competencies people use to make sense of the world, including but not limited to folk psychology. This provocative picture of the mind is supported by the fact that, originally thought to be restricted to the use of “intentionally” in the context of side-effects, Knobe’s results have been shown to extend to mental states such as “intending” and “knowing” (Beebe & Buckwalter, 2010; Pettit & Knobe, 2009), and to the means whereby an agent achieves her goal (Cova & Naar, 2012a).

### Two Methodological Challenges to the Knobe Effect

Because of their potential implications, accounts of the Knobe Effect in terms of moral judgments have been widely discussed and criticized (Cova, 2016a). However, one kind of criticism is particularly radical: It does not reject the idea that the Knobe Effect should be explained by appeal to moral evaluations, but denies the very existence of the Knobe Effect.

In its most basic and consensual formulation, the Knobe Effect can be described in the following terms: Because the difference between Knobe’s *Harm* and *Help* cases cannot be fully explained by the factors traditionally considered as relevant to intentionality judgments, then something else (maybe but not necessarily moral judgments) is also shaping these judgments. However, some have rejected this claim, and argued that traditional factors are in fact sufficient to explain people’s intentionality judgments. The fact that it does not seem so at first sight is only due to methodological issues supposed to plague Knobe’s studies and others designed on the same model.

There are two main methodological challenges to Knobe’s results: first, that people’s answers to Knobe’s questions might not reflect what they really think and, second, that people’s answers might be explained by the fact that they do not make the same assumptions about the agent’s mental states and attitudes in both cases. In this article, we address both methodological criticisms and investigate whether Knobe’s results survive when correcting for these presumed methodological shortcomings. Study 1 addresses the first challenge, while Studies 2 and 3 address the second.

### First Methodological Challenge: How to Accurately Measure Intentionality Judgments?

One criticism that has often been raised against Knobe’s results is that participants’ answers in these studies might not
reflect what they actually think. The first to raise such a criticism were Adams and Steadman (2004), who interpreted the Knobe Effect as the product of conversational implicatures. According to them, people use the word “intentionally” in the Harm case to conversationally imply that the agent deserves blame as they are offered no other way to express their moral disapproval. In consequence, Adams and Steadman (2007) tried to find a better way of measuring intentionality judgments. They gave participants the Harm case and then asked them to select the best answer among these two:

- “The chairman harmed the environment knowingly, but not intentionally.”
- “The chairman harmed the environment knowingly and intentionally.”

However, in line with Knobe’s results, 80% of participants still chose the second option.

Guglielmo and Malle (2010) recently raised a similar objection. According to them, participants’ answers in the Harm case cannot be taken at face value. Even if the chairman did not intentionally harm the environment, it is still true that he did many things intentionally (such as intentionally starting a program that would harm the environment). However, because Knobe’s forced-choice questions does not offer people any other way to express the fact that the relation between the chairman’s decision and the environmental harm is not accidental, participants have no other choice but to say that he intentionally harmed the environment.

To correct for this problem, Guglielmo and Malle (2010) gave participants cases similar to the Harm case and asked them to choose which description of the chairman’s action they found the most accurate among the following four: “The chairman willingly harmed the environment,” “The chairman knowingly harmed the environment,” “The chairman intentionally harmed the environment,” and “The chairman purposefully harmed the environment.” It turns out that very few participants chose the description according to which the chairman intentionally harmed the environment (see Table 1). This suggests that the Knobe Effect might indeed be a mere methodological artifact.

<table>
<thead>
<tr>
<th>Description</th>
<th>Most accurate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The CEO willingly harmed the environment”</td>
<td>12</td>
</tr>
<tr>
<td>“The CEO knowingly harmed the environment”</td>
<td>86</td>
</tr>
<tr>
<td>“The CEO intentionally harmed the environment”</td>
<td>1</td>
</tr>
<tr>
<td>“The CEO purposefully harmed the environment”</td>
<td>1</td>
</tr>
</tbody>
</table>

However, Guglielmo and Malle’s methods also have their shortcomings. First, their results are hard to interpret: It is not because participants consider a claim not to be the most accurate that they consider it to be false. For example, Cova (2016b) found that most participants still refused to choose the claim that “the chairman intentionally harmed the environment” in a case in which harming the environment constituted a means for the chairman’s action and was therefore clearly intended. Second, the method seems to yield inconsistent results: Cova (2016b) used Guglielmo and Malle’s method but asked participants to choose the least accurate statement. The statement that was less often chosen as “least adequate” was that “the chairman intentionally harmed the environment.”

Nevertheless, the fact that it proves hard to create new, reliable measures of intentionality judgments does not undermine the main methodological criticism that participants’ answers in Knobe’s studies might not reflect what they really think, and that they might say that the chairman intentionally harmed the environment only to express something else. Study 1 thus aims to determine whether people tend to judge the chairman’s action less intentional when they are offered (a) alternate possibilities to express their moral disapproval and (b) alternate ways of describing the chairman’s behavior.

**Study 1**

In the present study, we compare standard measures of intentionality (the yes/no question and the agreement scale) with new measures of intentionality that give more choice to participants. These new measures, inspired from Guglielmo and Malle’s (2010) work, present the advantage (a) of giving participants the opportunity to blame/praise the agent for her behavior and (b) of providing participants with alternate descriptions of the chairman’s behavior. Thus, these new measures should avoid the methodological objections that have been raised against standard measures of intentionality.

Our goal in this study was to determine whether standard measures of intentionality really tend to overestimate participants’ ascriptions of intentionality, when compared with these new measures of intentionality.

**Participants**

Participants were 240 individuals subscribed as workers on Amazon Mechanical Turk (MTurk) and located in the United States ($M_{\text{age}} = 36.09, SD_{\text{age}} = 11.42$; 47% male). They were recruited online and paid US$0.25 for their participation. Participants were assigned to one of the following conditions: Standard, Four items, or Five items.

**Materials and Procedure**

After providing basic demographic information, participants were presented with the following case:
Forest—The vice-president of a company went to the chairman of the board and said, “We are thinking of starting a new program. We are sure that it will help us increase profits. However, starting the program will require using new toxic chemicals and releasing them in the air, which will cause acid rains that will end up destroying the forest near the factory.”

The chairman of the board answered, “I don’t care at all about destroying the forest. 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 forest was destroyed.

After reading the text, participants in the Standard condition received the following question:

1. Did the chairman intentionally destroy the forest? (Yes/no/both/neither/I don’t know)
2. To what extent do you agree with the following statement: “The chairman intentionally destroyed the forest?” (1 = completely disagree, 7 = fully agree).

Participants in the Four items and Five items conditions received first the following question:

For each of the following statements, indicate whether it is a Correct or Incorrect description of the situation (correct/incorrect/both/neither/I don’t know; statements were presented in a random order):

1. The chairman destroyed the forest.
2. The chairman deserves blame for destroying the forest.
3. The chairman deserves praise for destroying the forest.

This question was presented first to make sure that participants had the occasion to express their moral disapproval of the chairman’s behavior and did not feel compelled to use the intentionality question to express it.

Participants in the Four items condition then received the following question:

For each of the following statements, indicate whether it is a Correct or Incorrect description of the chairman’s behavior (correct/incorrect/both/neither/I don’t know; statements were presented in a random order):

1. The chairman intentionally destroyed the forest.
2. The chairman willingly destroyed the forest.
3. The chairman knowingly destroyed the forest.

Participants in the Five items condition received the following question:

For each of the following statements, indicate whether it is a Correct or Incorrect description of the chairman’s behavior (correct/incorrect/both/neither/I don’t know; statements were presented in a random order):

1. The chairman intentionally destroyed the forest.
2. The chairman intentionally put profits before the forest.
3. The chairman intentionally adopted a program he knew would destroy the forest.
4. The chairman intentionally disregarded the forest when adopting the program.
5. The chairman intentionally started a forest-destroying program.

Finally, participants from all three conditions received two assumptions checks about the chairman’s intent and the nature of her action (goal/means/side-effect), the latter being

Which of the following statements most accurately describes the situation?

- Destroying the forest was the chairman’s goal.
- Destroying the forest was a means for the chairman to reach his goal.
- Destroying the forest was a side-effect of the chairman’s attempt to reach his goal.
- None of the above.

Results and Discussion

Assumption check. Because debates about the Knobe Effect are mainly focused on attributions of intentionality in the case of side-effects, we excluded from analyses all participants who did not answer that “destroying the forest was a side-effect of the chairman’s attempt to reach his goal” to the assumption check. We were left with 49 participants in the Standard condition, 56 in the Four items condition, and 57 in the Five items condition.

Intentionality judgments. Answers to the different measures of intentionality are presented in Table 2. There was no significant difference between the proportion of people endorsing the claim that the chairman intentionally destroyed the forest in the Standard and Four items conditions, \( \chi^2(1, N = 105) = 0.13, p = .72, Q = 0.06 \). Nor was there a significant difference between the Standard and Five items conditions, \( \chi^2(1, N = 106) = 0.41, p = .52, \Phi = 0.09 \). Thus, there is no reason to think that standard measures of intentionality overestimate participants’ ascriptions of intentionality. In fact, standard measures were those that returned the lowest ascriptions of intentionality.

| Table 2. Participants’ Ascriptions of Intentionality in Study 1 (After Exclusion). |
|-----------------|-----------------|-----------------|
|                | Standard        | Four items      | Five items      |
| Intentionality  | Yes: 75.5%      | Correct: 80.4%  | Correct: 82.5%  |
| No: 12.2%       | Incorrect: 14.3%| Incorrect: 7.0% |
| Scale: 5.99 (1.44) |                 |                 |

1. The chairman intentionally destroyed the forest.
2. The chairman intentionally put profits before the forest.
3. The chairman intentionally adopted a program he knew would destroy the forest.
4. The chairman intentionally disregarded the forest when adopting the program.
5. The chairman intentionally started a forest-destroying program.
Together, these results suggest that the high ascriptions of intentionality obtained using standard measures of intentionality actually reflect participants’ real judgments. Even when participants are given the chance to express their moral disapproval and provided with alternative descriptions, the majority of them continue to endorse the claim that the chairman intentionally brought about a negative side-effect.

Second Methodological Challenge: Controlling for Mental States Ascriptions

However, even if we can trust the standard measures of intentionality, we still have a second reason to doubt that there is such a thing as the Knobe Effect. Indeed, as we pointed out, it is commonly accepted that what makes Knobe’s initial results intriguing is that the asymmetry between the Harm and Help cases cannot be explained by the chairman’s attitudes toward the outcome of his actions as he is indifferent in both cases. However, recent results have challenged this key assumption by showing that participants attribute the chairman a desire to bring harm to the environment in the Harm case, but not a desire to help the environment in the Help case (Guglielmo & Malle, 2010; Pettit & Knobe, 2009; Uttich & Lombrozo, 2010).

Based on these results, some have argued that Knobe’s results can in fact be completely explained by traditional factors (such as the agent’s desire to bring about the outcome) and that there is no need to bring moral considerations into the equation. To support this rejection of the Knobe Effect, Guglielmo and Malle (2010) have shown that intentionality judgments can be lowered in the Harm case by having the chairman express regret, and increased in the Help case by having the chairman expressing his joy at helping the environment, to the point that intentionality ascriptions are lower in the Harm than in the Help case (see Figure 1). They end up concluding that “when controlling for desire strength, the effect of moral valence all but disappears” (Guglielmo & Malle, 2010, p. 1640).

However, this conclusion might be premature, for at least two reasons. The first is that, as far as moral valence is concerned, Guglielmo and Malle (2010) focused only on blame judgments. The second is that showing that the asymmetry between the Harm and the Help case can disappear when we compare a case of joyful help and a case of regretful harm says nothing about the role of normative evaluations as no one has ever argued that normative evaluations are the sole determinant of intentionality judgments, or that the asymmetry would hold no matter the agent’s attitudes. The problems raised by these two limitations will appear more clearly in light of the most influential account of the Knobe Effect: Knobe’s own account.

When it comes to explaining the asymmetry he discovered, Knobe (2010) considered that the key moral evaluations are participants’ evaluations of the outcome’s valence (as good or bad). According to Pettit and Knobe’s account, that we will call the “Normative Expectations Hypothesis” (Pettit & Knobe, 2009), a side-effect is intentional when the agent’s attitudes toward the outcome are beyond a certain default point. This default point is the attitude a moral agent should have in a given situation. In the Harm case, the chairman should be reluctant to harm the environment but, because he is indifferent, his attitudes are beyond this default point—his action is thus intentional. In the Help case, the chairman should desire to help the environment but, because he is indifferent, his attitudes are below this default point—his action is thus not intentional (see Figure 2).

Pettit and Knobe’s account makes room for an impact of both the agent’s attitudes and the participants’ moral evaluations of the outcome. Several accounts have followed them on this point (Cova, Dupoux, & Jacob, 2012; Cova & Naar, 2012b; Egré, 2014). Thus, showing that intentionality judgments can be modified by varying the agent’s attitudes toward the outcome does not immediately rule out the possibility that moral evaluations also play a role. To reach this conclusion, one should compare good and bad cases in which the agent’s desire to bring...
about the outcome is kept constant. In this case, standard accounts of intentional action should predict that intentionality judgments will not differ across cases, while the “Normative Expectations Hypothesis” predicts that intentionality judgments will still be higher in cases involving bad outcomes. Guglielmo and Malle’s Regretful Harm and Joyful Help cases cannot bear on this issue as desire attributions are largely lower in the first case than in the second (see Figure 1). Of more direct relevance is a comparison between their Joyful Help and Harm cases: While desire attributions are similar in both cases, intentionality ratings are much higher in the Harm case. This pattern suggests that the same amount of desire leads to higher intentionality judgments when the side-effect is something the agent should not desire to bring about—exactly what the Normative Expectations Hypothesis would predict.

It is thus not clear whether controlling for the agent’s attitudes (and particularly desire) toward the outcome makes the apparent effect of moral considerations disappear. Study 2 investigates whether taking into account the agent’s desires really prevent moral considerations from mediating (and thereby explaining) the impact of bad versus good outcomes on intentionality judgments.

**Study 2**

One way to test for the respective impact of agent’s attitudes and participant’s moral expectations on ascriptions of intentionality (and intent) is to vary both factors independently, and thus to have four scenarios: one case in which the agent joyfully harms the environment, one case in which the agent reluctantly harms the environment, one case in which the agent joyfully helps the environment, and one case in which the agent reluctantly helps the environment. Most investigations of the Knobe Effect have only compared the Harm and Help cases, which vary both factors simultaneously, thereby preventing researchers to properly determine the respective role of each factor in shaping intentionality judgments (Sripada & Konrath, 2011). We thus decided to create new cases that would put a greater emphasis than the original Harm case on the chairman’s attitudes toward the environment:

**Joyful Harm**—The vice-president of a company went to the chairman of the board and said, “We are thinking of starting a new program. We are sure that it will help us increase profits. However, starting the program will require releasing new toxic chemicals in the air, which will cause acid rains that will end up destroying the forest near the factory. I know you really hate this forest, because you often went hiking there, and it is the place where you met and proposed your wife. Thus, I knew you would be upset to learn that the program would also harm the forest.”

The chairman of the board answered, “You’re right. This is indeed very upsetting news to me. Because this was not my purpose, I’m very sad to learn that the program will also harm the forest! However, it’s crucial that we increase profits. Let’s start the new program and use these new chemicals.”

They started the new program. Sure enough, the forest was destroyed.

**Regretful Harm**—The vice-president of a company went to the chairman of the board and said, “We are thinking of starting a new program. We are sure that it will help us increase profits. However, starting the program will require releasing new toxic chemicals in the air, which will cause acid rains that will end up destroying the forest near the factory. I know you really like this forest, because you often went hiking there, and it is the place where you met and proposed your wife. Thus, I knew you would be upset to learn that the program would also harm the forest.”

The chairman of the board answered, “You’re right. This is indeed very upsetting news to me. Because this was not my purpose, I’m very sad to learn that the program will also harm the forest! However, it’s crucial that we increase profits. Let’s start the new program and use these new chemicals.”

They started the new program. Sure enough, the forest was destroyed.

**Joyful Help**—The vice-president of a company went to the chairman of the board and said, “We are thinking of starting a new program. We are sure that it will help us increase profits. Moreover, starting the program will cause the release of new organic fertilizers in the air, which will cause the forest near the factory to grow and bloom. I know you really like this forest, because you often went hiking there, and it is the place where you met and proposed your wife. Thus, I knew you would be happy to learn that the program would also help the forest.”

The chairman of the board answered, “You’re right. This is indeed very good news to me. Though this was not my purpose, I’m thrilled to learn that the program will also help the forest! And it’s crucial that we increase profits. Let’s start the new program and use these new chemicals.”

They started the new program. Sure enough, the forest grew and bloomed.

**Regretful Help**—The vice-president of a company went to the chairman of the board and said, “We are thinking of starting a new program. We are sure that it will help us increase profits. Moreover, starting the program will cause the release of new organic fertilizers in the air, which will cause the forest near the factory to grow and bloom. I know you really hate this forest, because your father often went hiking there, and it is the place where he died from a hunting accident. Thus, I knew you would be upset to learn that the program would also help the environment.”

The chairman of the board answered, “You’re right. This is indeed very sad news to me. Because this was not my purpose, I’m very sad to learn that the program will also help the forest! However, it’s crucial that we increase profits. Let’s start the new program and use these new chemicals.”

They started the new program. Sure enough, the forest grew and bloomed.
The four cases differ from one another along two factors: the nature of the side-effect (Harm/Help) and the agent’s attitudes toward the side-effect (Joyful/Regretful). The predictions made by the Normative Expectations Hypothesis for these cases are the following. First, it predicts the Help/Harm factor should have a significant effect on ascriptions of intentionality, even if our design allows to vary the nature of the side-effect and the agent’s attitudes independently from one another. Second, it predicts that this effect should be mediated by participants’ normative expectations, even when taking into account the desires participants ascribe to the agent.

Participants

Participants were 400 individuals located in the United States and subscribed as workers on MTurk (\(M_{age} = 34.07, SD_{age} = 11.81\); 58% male). They were recruited online and paid US$0.3 for their participation.

Materials and Procedure

Participants received either the Joyful Harm, Regretful Harm, Joyful Help, or Regretful Help case. Then, participants were asked to rate their agreement with the five following claims on a scale (1 = completely disagree, 4 = neither agree, nor disagree, 7 = strongly agree; all claims were presented in a random order):

1. “The chairman intentionally harmed [helped] the forest.”
2. “The chairman wanted to harm [help] the forest.”
3. “The chairman was reluctant to harm [help] the forest.”
4. “The chairman should have wanted to harm [help] the forest.”
5. “The chairman should have been reluctant to harm [help] the forest.”

Then, participants were asked the following questions:

6. To what extent do you agree with the following statement: “The chairman intended to harm [help] the forest” (on a 7-point scale ranging from 1 = strongly disagree to 7 = strongly agree).
7. Which of the following statements most accurately describes the situation?

Answers to Question 1 constituted the intentionality score. Answers to Question 3 were reverse-scored and averaged with answers to Question 2 to constitute the desire score (\(r = .64, p < .001\)). Answers to Question 5 were reverse-scored and averaged with answers to Question 4 to constitute the normative expectations score (\(r = .77, p < .001\)).

Results

Assumption check. As in Study 1, we began by excluding from analyses all participants who did not answer that “harming [helping] the forest was a side-effect of the chairman’s attempt to reach his goal” to the last question. We were left with a total of 301 participants (75% of the original sample), including 73 participants for the Joyful Harm case, 65 in the Regretful Harm case, 84 in the Joyful Help case, and 79 in the Regretful Help case. The percentage of participants who were rejected differed significantly between conditions, \(\chi^2(3, N = 400) = 10.78, p = .012, V = 0.16\). Overall, participants were more likely to give another answer when the action was bad (harming). This is consistent with previous findings suggesting that moral considerations can affect people’s classification of actions as goals, means, or side-effects (Knobe, 2010; Ulatowski, 2012). Results for each score for each case are presented in Table 3 and the bivariate correlation between all the main variables are presented in Table 4.

| Table 3. Means (and Standard Deviations) for Each Score in Study 2, as a Function of Case (\(N = 301\)). |
|-----------------------------------------------|-----------------|-----------------|-----------------|-----------------|
| Intentionality                               | Joyful harm     | Regretful harm  | Joyful help     | Regretful help  |
|                                              | 5.62 (1.42)\(^a\) | 4.94 (1.75)\(^b\) | 3.20 (1.75)\(^c\) | 2.62 (1.78)\(^c\) |
| Desire scores                                | 6.11 (0.98)\(^a\) | 3.92 (1.22)\(^c\) | 5.53 (1.06)\(^b\) | 2.30 (1.26)\(^d\) |
| Normative expectations score                 | 1.95 (0.95)\(^c\) | 1.94 (1.07)\(^c\) | 5.74 (0.96)\(^a\) | 5.12 (1.52)\(^b\) |

Note. The difference between cases was tested using Tukey’s post hoc tests. In rows, means with different superscripts are significantly different (\(p < .05\)). See Supplementary Materials for \(p\) values and confidence intervals.

| Table 4. Bivariate Correlations Between All the Main Variables (Study 2; \(N = 301\)). |
|-----------------------------------------------|-----------------|-----------------|
| 1. Intentionality                           | 2. Desire       | 3. Normative expectations |
| 1. —                                          | —               | —               |
| 2. .47***                                     | —               | —               |
| 3. .55***                                     | −.18***         | —               |
| \(M\) 4.01                                    | 4.47            | 3.84            |
| \(SD\) 2.08                                   | 1.88            | 2.10            |

*** \(p < .001\).
Intentionality judgments. To test our first prediction, we ran a two-way ANOVA to test the effect of the Joyful/Regretful factor and the Harm/Help factor on the intentionality judgments. Results revealed a significant main effect of the Joyful/Regretful factor indicating higher judgments of intentionality for the agent who joyfully acted on the environment ($M = 4.32$, $SD = 2.01$, $n = 157$) than for the agent who regretfully acted on the environment ($M = 3.67$, $SD = 2.11$, $n = 144$), $F(1, 297) = 10.43$, mean square error ($MSE = 843.20$, $p = .001$, $η_p^2 = .034$, 95% confidence interval (CI) = $[-1.01, -0.25]$. The main effect of the Harm/Help factor indicates higher judgments of intentionality for the agent who harmed the environment ($M = 5.30$, $SD = 1.61$, $n = 138$) than for the agent who helped the environment ($M = 2.92$, $SD = 1.78$, $n = 163$), $F(1, 297) = 147.04$, $MSE = 843.20$, $p < .001$, $η_p^2 = .331$, 95% CI = $[1.98, 2.75]$. The interaction of these two factors was not significant, $F(1, 297) < 1$.

Mediation analysis. To test our second prediction, that is, to determine whether normative expectations play a unique role (beyond the role of desire) in shaping intentionality judgments, we ran a multiple mediation model by following Preacher and Hayes’s (2008) bootstrapping procedure (percentile bootstrap procedure with 5,000 bootstrap samples), using the SPSS version of the Preacher and Hayes (2008) macro (Model 4). Results of these analyses suggest that normative expectations and desire significantly mediated the relationship between the Harm/Help condition (collapsing the two Harm cases: Joyful Harm and Regretful Harm, and the two Help cases: Joyful Help and Regretful Help) and intentionality judgments (see Figure 3; the two indirect effects are significant at $p < .001$ and are presented in detail in Table 5). Hence, in comparison with the agent who helped the environment, the agent who harmed the environment was seen as having more desire to harm the environment, which in turn increases intentionality judgments beyond the effect of normative expectations. Moreover, in comparison with the agent who helped the environment, the agent who harmed the environment was less expected to desire to affect the environment, which in turn increased intentionality judgments beyond the effect of desire. The direct effects remain significant after taking into account the two abovementioned mediation effects, $B(297) = 0.46$, $p = .006$.

Discussion

For intentionality judgments, our results confirmed the two predictions of the Normative Expectation Hypothesis: (a) manipulation of the nature of the side-effect (Harm/Help) had a significant effect on intentionality judgments, even in a design that made it vary orthogonally to the agents’ attitudes.
(Joyful/Regretful), and (b) this effect was mediated by participant’s normative expectations, even when taking into account the desires participants attributed to the agent.

These results suggest that the impact of the Harm/Help factor (i.e., the Knobe Effect) cannot be explained by the sole attitudes participants attribute to the agent: As predicted by the Normative Expectations Hypothesis, normative expectations also play a role. Moreover, our results even support more specific predictions of the Normative Expectations Hypothesis. As mentioned earlier, the Normative Expectations Hypothesis predicts that lower amounts of desire will be required for an action to count as intentional when the outcome is something we expect the agent not to desire. This is exactly what a direct comparison of the Joyful Help and Regretful Harm cases suggest (see Table 3). While desire ratings are lower in the Regretful Harm case ($M = 3.92$, $SD = 1.22$) than in the Joyful Help case ($M = 5.53$, $SD = 1.06$), $t(147) = 8.61$, $p < .001$, $d = 1.42$, $95\%$ CI $= [0.62, 0.99]$, intentionality judgments are higher in the first case than in the second one, $t(147) = 6.00$, $p < .001$, $d = 0.99$, $95\%$ CI $= [-1.15, -0.58]$. The fact that desires and intentionality judgments exhibit reversed patterns in these two cases goes directly against the idea that intentionality judgments could be explained only in terms of the attitudes participants attribute to the agent (see Figure 4).

Overall, our results suggest that, even if agents’ attitudes (desires) obviously have an effect on intentionality judgments, this does not prevent normative expectations to shape those judgments as well. The data are more consistent with an account of intentionality judgments that makes room for both factors, than with a single-factor account that would not leave room for moral considerations.

### Study 3

Study 2’s results suggest that agent’s attitudes are not enough to account for participants’ intentionality judgments and that there is an irreducibly normative factor driving them. However, we have not yet exhausted all the accounts that have been put forward to explain away the Knobe Effect. Indeed, the chairman’s desire to bring about the relevant outcome might not be the only attitude to change between the Help and Harm case. Other attitudes have been hypothesized to vary between the two cases, and thus, to explain the asymmetry in intentionality judgments:

**Deep Self:** One potential shortcoming of Study 2 is that we might not have focused on the relevant type of attitudes by only asking participants about what the chairman wanted. According to Sripada’s “deep self model” of intentionality judgments, what drives participants’ attributions of intentionality is not whether the agent actually desires to bring about the outcome or is reluctant to in the moment, but whether this action matches her deep-held values and principles (what she truly desires, deep in herself, Sripada, 2010, 2012; Sripada & Konrath, 2011).

**Taking Into Consideration:** Another account of the Knobe Effect emphasizes the extent to which the chairman takes the side-effect into consideration when making his decision. According to Scaife and Webber’s (2013) “Consideration Hypothesis” (see also Cova, 2014), a side-effect is judged intentional when participants consider that the agent gave it some consideration and paid it some attention when deliberating about her action.

**Reasons Against:** According to Frank Hindriks (2008, 2014), a side-effect is judged intentional when the agent carelessly brings about an outcome she thought counted as a reason not to act (see Cova, 2016c, for a discussion).

**Knowledge and Belief:** Finally, it has recently been found that the Knobe Effect extended to epistemic concepts, such as knowledge or belief: People are more likely to say that the chairman knew or believed that starting the new program would bring about the relevant side-effect when this side-effect is bad (Beebe, 2013; Beebe & Buckwalter, 2010). Because one needs to believe that one’s action will bring about a given outcome to intentionally bring it about, it has been proposed that the asymmetry in belief attributions was the source of the asymmetry in intentionality judgments (Alfano, Beebe, & Robinson, 2012).

The goal of the present study is to test whether normative expectations still mediates the effect of the nature of the side-effect (good or bad) on intentionality judgments, even when taking into account a whole range of attitudes. Moreover, Study 2 was limited by the fact that we failed to perfectly match desire ascriptions between Harm and Help cases, and thus to vary both factors independently. In both Joyful and

---

**Figure 4.** Intentionality scores and desire scores (on a scale from 1 to 7) for the Joyful Harm, Regretful Harm, Joyful Help, and Regretful Help cases in Study 2.

Note. Results of two t tests show that, though desire scores are higher in the Joyful Help case than in the Regretful Harm cases, intentionality scores are still lower in the Joyful Help case. The error bars represent the standard errors of the mean.
Regretful conditions, participants attributed a greater desire to bring about the outcome to the agent in Harm cases. Thus, we designed a new series of cases, replacing the Joyful Harm and Regretful Help cases by cases in which the agent is indifferent to the side-effect. Our hope was that this new series of cases would prove more successful in matching the attitudes participants attribute to the agent between the Harm and Help conditions:

Indifferent Harm (replaces Joyful Harm)—The vice-president of a company went to the chairman of the board and said, “We have finally come up with a new program. We are sure that it will help us increase profits and allow us to save our company. Also, I know you told us not to care about the potential environmental side-effects starting, but I thought you might be interested to know that starting the program will require releasing new toxic chemicals in the air, which will cause acid rains that will end up destroying the forest near the factory.”

The chairman of the board answered, “And why would I be interested in knowing such a thing? Don’t care at all whether our program ends up harming or helping the environment. I just want to make as much profit as I can and save my company. Let’s start the new program.”

They started the new program. Sure enough, the forest was destroyed and the environment was harmed.

Indifferent Help (replaces Regretful Help)—The vice-president of a company went to the chairman of the board and said, “We have finally come up with a new program. We are sure that it will help us increase profits and allow us to save our company. Also, I know you told us not to care about the potential environmental side-effects, but I thought you might be interested to know that starting the program will cause the release of new organic fertilizers in the air, which will cause acid rains that will end up destroying the forest near the factory and bloom.”

The chairman of the board answered, “And why would I be interested in knowing such a thing? Don’t care at all whether our program ends up harming or helping the environment. I just want to make as much profit as I can and save my company. Let’s start the new program.”

They started the new program. Sure enough, the forest grew and bloomed and the environment was helped.

The two remaining cases were also modified accordingly (vignettes’ full text can be found in Section 1 of Supplementary Materials).

Participants

Participants were 240 individuals located in the United States and subscribed as workers on MTurk ($M_{age} = 33.95, SD_{age} = 10.98; 50\%$ male). They were recruited online and paid US$0.4 for their participation.

Materials and Procedure

Participants received either the Indifferent Harm, Regretful Harm, Joyful Help, or Indifferent Help case ($n = 60$ by condition). Then, participants answered 11 questions on a seven-item scale (all questions were presented in a random order, one question per page). More precisely, intentionality was measured by asking participants to what extent they agreed with the statement, “The chairman intentionally harmed [helped] the environment” (–3 = strongly disagree, 3 = strongly agree). The deep self score was measured by asking participants to what extent they agreed with the statements, “The chairman truly desired to harm [help] the environment” and “Deep down, the chairman did not really want to harm [help] the environment” (the last statement was reverse-scored; −3 = strongly disagree, 3 = strongly agree; $r = .50, p < .001$). The consideration score was measured by asking participants, “In making his decision, to what extent did the chairman take into consideration the fact that starting the new program would harm [help] the environment?” and “To what extent did the chairman pay attention to the fact that starting the new program would harm [help] the environment?” (−3 = not at all, 3 = very much; $r = .70, p < .001$). The reason-against score was measured by asking participants, “To what extent did the chairman believe that harming [helping] the environment counted as a reason NOT TO start the new program?” and “To what extent did the chairman believe that he should NOT start the new program because it would harm [help] the environment?” (−3 = not at all, 3 = very much; $r = .52, p < .001$). To form the belief score we averaged the statements, “After the discussion with the vice-president, when he took the decision to start the new program, did the chairman know that the new program would harm [help] the environment?” (−3 = the chairman did not know, 3 = the chairman knew) and “After the discussion with the vice-president, when he took the decision to start the new program, did the chairman believe that the new program would harm [help] the environment?” (1 = the chairman did not believe, 3 = the chairman believed; $r = .47, p < .001$). Finally, the normative expectations score was computed by averaging the statements “To what extent should the chairman have desired to harm [help] the environment?” and “To what extent should the chairman have been reluctant to harm [help] the environment?” (the second statement was reverse-scored, −3 = not at all, 3 = very much; $r = .66, p < .001$).

Results

Results for each score for each case are presented in Table 6 and the bivariate correlation between all the measures variables are presented in Table 7.

Intentionality judgments. We ran a one-way ANOVA to assess the effect of the different cases (Indifferent Harm, Regretful
Harm, Joyful Help, or Indifferent Help) on the intentionality judgments, which reveals significant differences between conditions, $F(3, 236) = 83.90, MSE = 2.60, p < .001, \eta^2_G = 0.52$. Post hoc (Tukey) tests reveal only one case that differs from the others: Intentionality judgments are definitely lower in the Indifferent Help case than all the other cases, the latter being not significantly different from each other (see Table 6 for the different values).

Mediation analysis. To determine the role of a larger range of attitudes in shaping intentionality judgments, we ran a multiple mediation model by following Preacher and Hayes’s (2008) bootstrapping procedure (percentile bootstrap procedure with 5,000 bootstrap samples). Results of these analyses suggest that, out of five possible mediating variables, only normative expectations and beliefs significantly mediated the relationship between the Harm/Help condition (again by collapsing the two Harm cases: Indifferent Harm and Regretful Harm, and the two Help cases: Joyful Help and Indifferent Help) and intentionality judgments (see Figure 5; the two indirect effects are respectively significant at $p < .001$ and are presented in detail in Table 8). An agent who harmed the environment is seen as having stronger beliefs in the consequence of her actions than an agent who helped the environment, which in turn increases intentionality judgments, above and beyond the effect of other variables. Finally, the direct effect is still significant when all the other variables (including the two mediation effects) are taken into account, $B(233) = 0.60, p < .001$. Thus, it seems that normative expectations do explain part of the Knobe Effect, even when controlling for a whole range of the agent’s attitudes.

General Discussion

In this article, our goal was to determine whether the claim that moral evaluations have an impact on intentionality judgments could survive two methodological challenges. The first challenge was the suggestion that high intentionality judgments in the Harm case might only be a methodological artifact and that people might endorse the claim that the chairman intentionally harmed the environment only to express something else. Study 1 investigated this claim by comparing standard measures of intentionality with new measures that gave participants more opportunity to express and endorse alternate statements. However, using these new measures did not decrease intentionality judgments, suggesting that standard measures of intentionality do not overestimate participants’ judgments.

The second challenge was the hypothesis that the asymmetry between the Harm and Help cases can be completely explained by appeal to non-moral factors, such as the agent’s

| Table 6. Means (and Standard Deviations) for Each Score in Study 3, as a Function of Case ($N = 240$). |
|-----------------------------------------------|----------------|----------------|----------------|----------------|
| Intentionality score | 1.85 (1.66)a | 1.72 (1.81)a | 1.60 (1.40)a | −2.08 (1.54)b |
| Deep self score | 0.92 (1.32)b | −0.40 (1.58)c | 1.73 (1.39)a | −1.48 (1.34)d |
| Consideration score | −0.98 (2.03)c | 0.18 (1.75)b | 1.98 (1.18)a | −1.87 (1.46)d |
| Reason-against score | −2.10 (1.50)b | −1.19 (1.77)a | −2.40 (1.13)b | −2.19 (1.12)b |
| Belief score | 2.41 (0.96)a | 2.38 (1.22)a | 2.06 (1.42)a | 1.18 (1.63)b |
| Normative expectations score | −1.66 (1.68)b | −1.68 (1.52)b | 2.08 (1.18)a | 1.55 (1.38)b |

Note. The difference between cases was tested using Tukey’s post hoc tests. In rows, means with different superscripts are significantly different ($p < .05$). See Supplementary Materials for $p$ values and confidence intervals.

| Table 7. Bivariate Correlations Between All the Measured Variables (Study 3, $N = 240$). |
|-----------------------------------------------|---------------|---------------|---------------|---------------|
| 1. — & — & — & — & — & — |
| 2. .55*** & — & — & — & — & — |
| 3. .39*** & .35*** & — & — & — & — |
| 4. .05 & −.23*** & .10 & — & — & — |
| 5. .43*** & .24*** & .27*** & −.19*** & — & — |
| 6. −.33*** & .09 & .19*** & −.26*** & −.18*** & — |
| M | 0.77 | 0.19 | −0.17 | −1.97 | 2.00 | 0.07 |
| SD | 2.30 | 1.86 | 2.17 | 1.47 | 1.41 | 2.27 |

*p < .01, ***p < .001.
desire to bring about the outcome. To test for this claim, Study 2 focused on a particular kind of moral evaluations, normative expectations about the agent’s attitudes. Results showed that, even when controlling for the agent’s desire to bring about the outcome, normative expectations mediated the impact of the side-effect’s nature (good or bad) on intentionality judgments. Study 3 showed that this was still the case, even when controlling for a wider range of attitudes.

Thus, it seems that moral evaluations still play an irreducible role in shaping our judgments of intentionality. However, it does not mean, as we saw, that they are the only factors playing a role. Most accounts of the Knobe Effect have emphasized one factor because their aim was to reject the claim that moral evaluations had an impact on our attributions of intentionality. But the results of Studies 2 and 3 suggest that many factors have an influence on these attributions and contribute to explain the difference between the Help and Harm cases: not only moral evaluations about the attitudes the agent should hold but also judgments about her desire to bring about the side-effect and whether she actually believed that the action would produce the relevant side-effect. Thus, an integrative account of the Knobe Effect should not limit itself to one or two factors (contrary to what has been done so far) but should make room for all these factors.

In conclusion, the studies presented in this article suggest that the Knobe Effect is a real phenomenon that does not have a simple and straightforward explanation and that judgments of intentionality cannot be simply explained by the agent’s attitudes, but are also driven by moral considerations. What the correct account should look like, however, is still a riddle.

**Authors’ Note**

Anthony Lantian and Jordane Boudesseul contributed equally to this work.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Center of Competence in Research (NCCR) in Affective Sciences financed by the Swiss National Science Foundation (n° 51NF40-104897) and hosted by the University of Geneva.

Supplemental Material
The online supplemental material is available at http://psb.sagepub.com/supplemental.

References


