Review of Philosophy and Psychology (2023) 14:1443–1470
https://doi.org/10.1007/s13164-022-00651-z

# The Doxastic Heuristic and the Consequence Account of the Epistemic Side-Effect Effect

**Katarzyna Paprzycka-Hausman[1] · Bartosz Maćkiewicz[1] · Katarzyna Kuś[1] · Marta Zaręba[1]**

Accepted: 25 June 2022 / Published online: 28 July 2022
© The Author(s) 2022

## Abstract

We discuss two philosophical explanations of the epistemic side-effect effect: the doxastic heuristic account (Alfano et al. *The Monist 95 (2):* 264–289, 2012) and the consequence account (Paprzycka-Hausman Synthese 197: 5457–5490, 2020). We argue that the doxastic heuristic account has problems with explaining knowledge attributions in cases where the probability that the side effect will occur is low and where the side effect does not ultimately occur. It can explain why there is a difference between the harm and the help cases but it cannot explain why people are willing to attribute knowledge in the harm cases. Such attributions can be explained on the consequence account, which takes knowledge attributions in norm-violation cases to be due to the increased salience of a consequence-awareness claim (knowledge that a possible consequence of the chairman's action is that the environment would be harmed). We report the results of a new study that tests the predictions of both accounts. In some conditions, people attribute knowledge of the side effect even in cases where the chairman does not have the relevant belief. This result directly contradicts the central tenet of the doxastic heuristic account. Linear regression models of knowledge attribution that correspond to the two accounts were compared. The addition of different justification options significantly contributes to the predictive power of the statistical model. The consequence account can explain the pattern of justifications better than the doxastic heuristic account. Our findings support the consequence account and pose a challenge to the proponents of the doxastic heuristic account.

✉ Katarzyna Paprzycka-Hausman
kpaprzycka@uw.edu.pl

Bartosz Maćkiewicz
b.mackiewicz@uw.edu.pl

Katarzyna Kuś
kkus@uw.edu.pl

Marta Zaręba
zareba.ma@uw.edu.pl

[1] Faculty of Philosophy, University of Warsaw, Warsaw, Poland

Although the sturdiness of the Knobe effect (Knobe 2003a, b, 2004, 2007, 2010; Knobe and Mendlow 2004; Pettit and Knobe 2009) has been surprising, what has proven perhaps even more surprising is the robustness of the epistemic side-effect effect (ESEE). ESEE involves such concepts as knowledge (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2016), belief (Beebe 2013) and probability (Dalbauer and Hergovich 2013). It persists in a variety of conditions that challenge the classical concept of knowledge (Turri 2014; Beebe and Jensen 2012; Paprzycka-Hausman 2020).

In this paper, we discuss two philosophical explanations of ESEE: the doxastic heuristic account (Alfano et al. 2012; Robinson et al. 2015) and the consequence account (Paprzycka-Hausman 2020). After presenting ESEE (§1), we argue that the doxastic heuristic account (§2) has problems with accommodating some of the available evidence, which can be explained by the consequence account (§3). In §4, we present the results of a study designed to adjudicate between the two accounts.

## 1 The Epistemic Side-Effect Effects

In Knobe's original (2003a) study, people were presented with one of two stories (Harm and Help) that differed in the four marked places:

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

The subjects were asked whether they agreed or disagreed with the claim that the chairman's action was intentional. Most (82%) subjects agreed with the claim that the chairman intentionally harmed the environment, while most (78%) disagreed with the claim that he intentionally helped the environment. This asymmetry is known as "the Knobe effect".

Beebe and Buckwalter (2010) have found a similar effect for knowledge attributions (ESEE). In their study with Knobe's original scenarios, people were asked whether they agree that the chairman knew that the new program would harm/help the environment. In the harm case, the mean response was 2.25 while it was 0.91 in the help case (Likert scale between -3 and 3 was used). In a forced-choice study, Beebe and Jensen (2012) obtained nearly the classical Knobe-effect proportions: 68% of people attributed knowledge in the harm case but only 16% in the help case.

Both the Knobe effect and ESEE are surprising. The stories appear to be symmetrical with respect to the mental states of the chairman relevant to the attribution of intentional action. They also appear to be symmetrical with respect to justification, truth and belief, which are relevant to the attribution of knowledge. It is thus puzzling why there should be an asymmetry in the attribution of knowledge between the help and the harm scenarios. Unlike the Knobe effect, ESEE was not foreshadowed

by philosophical discussions.[1] It thus appears to be even more surprising and a genuine experimental philosophical discovery.

ESEE was replicated in other languages (Dalbauer and Hergovich 2013). It was also replicated on the concept of belief (Beebe 2013) as well as some concepts of probability (Dalbauer and Hergovich 2013). Beebe and Shea (2013) have shown that, in stories where some norm is violated ("harm" cases), the knowledge claim tends to persist despite the introduction of Gettier-type luck elements into the stories (see also Buckwalter 2014; Turri 2014; Yuan and Kim 2021). Moreover, ESEE can be found even in slight-chance of harm and in Butler-type scenarios (Beebe and Jensen 2012; Paprzycka-Hausman 2020), where attributions of knowledge seem to be prima facie unjustified because the probability that the effect will occur is small. Turri (2014) has demonstrated that people attribute knowledge in the harm scenario even when they are told that the environment was not harmed, which challenges the facticity of knowledge.

## 2 Doxastic Heuristic Account

The doxastic heuristic account (DHA) was proposed by Alfano et al. (2012) as the most comprehensive explanation of many side-effect effect studies, including ESEE (cf. Hindriks 2019). On their view, the side-effect effect for beliefs (Beebe 2013) is the basic side-effect effect, from which all the others derive.

The core thought of DHA is that, in cases of norm violation, there are additional reasons for people to reflect on the side effect and to form beliefs about it. It is rational for agents to be more epistemically alert and to deliberate more in cases of norm violation than in cases of norm conformity. For this reason, Alfano et al. argue that (a) when somebody else's action would lead to a result that violates a norm, we should attribute to that person the belief that the action would lead to that result (norm-violation/belief-attribution heuristic), and (b) when our action would lead to a result that violates a norm, we should form the belief that the action would lead to that result (norm-violation/belief-formation heuristic).

By appealing to these two central doxastic heuristics, they explain the side-effect effect for beliefs. Since it is rational to deliberate in cases of norm violation and part of the deliberation ought to involve considering side effects (see also Bratman 1987; Hindriks 2008), people will exhibit a greater tendency to form beliefs and consequently also to attribute beliefs in cases of norm violation than in cases of norm conformity (provided that people actually do what is rational).

In order to account for the other side-effect effects, Alfano et al. show that a properly reconstructed belief is plausibly regarded as a necessary condition for many mental state concepts. Indeed, they show that belief turns out to be a necessary

[1] Harman (1976) was the first to formulate exceptions to the simple view of intentional action. He allowed that intentionality be attributed in the absence of intention when there are reasons against the action. Both the Knobe effect and the Butler (1978) problem were foreshadowed by Harman's account.

condition not only for mind-to-world concepts (e.g. knowledge[2]) but also for some world-to-mind concepts (e.g. intentionality, intention, desire, acting in order to, etc.). Here are three examples of such belief principles for knowledge, desire, and intentionality:

(Know→Believe) Agent α knows that φ'ing would make it the case that *p* by φ'ing only if α believes that φ'ing would help to make it the case that *p*.
(Desire→Believe) Agent α desires to make it the case that *p* by φ'ing only if α believes that φ'ing would help to make it the case that *p*.
(Intentionally→Believe) Agent α intentionally makes it the case that *p* by φ'ing only if α believes that φ'ing would help to make it the case that *p*.

The general explanatory structure is the following. We begin with the epistemic side-effect effect for beliefs: people tend to attribute beliefs in norm-violation but not in norm-conformity cases. We can then use the belief principles (and something like *modus tollens*) to show that, in the norm-conformity cases, people will be less inclined to attribute other psychological states: knowledge, intentionality, desire, etc. In brief, since there is no belief, there is no knowledge, no intentionality, no desire, etc. (p. 268). DHA thus explains why people are less inclined to attribute knowledge, intentionality, desire, etc. in norm-conformity cases.

However, in some ESEE studies (false-harm and slight-chance-of-harm cases), what is puzzling is not only that (1) there is a difference between norm-violation and norm-conformity cases but also that (2) people tend to attribute knowledge in the norm-violation cases at all. We will argue that DHA can explain (1) but it cannot explain (2),[3] unless it appeals to a non-standard concept of knowledge.

In the false harm/help condition (Turri 2014, experiment I), after the chairman is told that the program would harm/help the environment, it turns out that it does not. In the false-harm condition, people's attributions of knowledge seem to be unshaken, despite the fact that the truth condition is not satisfied (false harm: M=4.76, on a 1–5 Likert scale). In the false-help condition, people hesitate whether to attribute knowledge (M=3.0). Arguably, DHA explains why there is a difference between the false-harm and the false-help condition. However, in this case, what really calls for an explanation is why people attribute knowledge in the false-harm case. DHA does not offer any tools to explain that.

A similar situation arises for the slight-chance of harm conditions, in which the justification condition (classically conceived) is not satisfied (Turri 2014; Beebe and

[2] The connection between knowledge and belief has been challenged in experimental research (e.g. Myers-Schulz and Schwitzgebel 2013; Murray et al. 2013). For a recent defence of the connection, both on conceptual and experimental grounds, see Buckwalter et al. (2015).

[3] DHA was later complemented with the *salient norm hypothesis* (Robinson et al. 2015), according to which the side-effect effects arise not simply when a norm is violated but when the violated norm is salient. While the authors present the hypothesis as a development of DHA, it is rather clear that the hypothesis can be accepted on other accounts that take norm-violation as a factor that is relevant in the explanation. In particular, there is no reason why such an additional hypothesis could not be adopted by the account discussed in §3. The hypothesis requires further study, however. The main problem is that, contrary to the authors' intentions, the two studies that were supposed to support *salient norm hypothesis* did not examine attributions for side effects but for main effects. It may be that the result could be obtained for side effects as well but this has not yet been shown. Second, it is not entirely clear whether the norm should be salient to the agent or to the attributor of the mental state in question.

Jensen 2012; Paprzycka-Hausman 2020). Beebe and Jensen (2012) were the first to find that people are inclined to attribute knowledge to the chairman even when the vice-president claims that there is a slight chance that the program will harm the environment (the mean was 1.15 in comparison to the mean of 2.25 in the original Beebe and Buckwalter 2010 study). In a study using slight-chance vignettes and a forced-choice paradigm (Paprzycka-Hausman 2020, study 4), 82.9% people attributed knowledge in the harm case while 56.2% did so in the help case. To the extent that DHA accepts that justification is a necessary condition on knowledge, it does not account for the slight-chance of harm experiments.[4] DHA explains why there is a difference in the harm and the help case in slight-chance scenarios but it does not explain why knowledge is attributed in the slight-chance of harm case.

Turri (2014) also claims that ESEE occurs in the absence of belief. This result might be taken to falsify the central tenet of DHA ("no belief, so no knowledge") since knowledge was attributed in the harm case. However, DHA theorists can plausibly challenge Turri's operationalization of the concept of belief: the belief that *p* was operationalized solely in terms of an explicit statement that one accepts that *p*. One can explicitly say that one accepts/rejects that *p* and really believe something else.

In sum, while DHA can explain why people are less inclined to attribute knowledge (and other mental states) in the norm-conformity cases, it does not have sufficient resources to explain why people tend to attribute knowledge in the norm-violation cases. In the case of the original ESEE, one can appeal to the fact that all the conditions of the JTB account of knowledge are satisfied to explain the knowledge attributions. The false-harm and the slight-chance of harm experiments, however, cannot be explained in such a way. Still, it should be stressed that DHA does offer a possible explanation why there is an asymmetry in all the cases. DHA could thus be thought of as offering a partial explanation, which might be adopted with some other explanation to provide a more complete understanding of the phenomena in question.

## 3 The Consequence Account of ESEE

Paprzycka-Hausman (2020) offers an account of ESEE that can explain attributions of knowledge in the slight-chance of harm and the false-harm scenarios. She argues that the culprit knowledge claim could be read as an abbreviation of a consequence-awareness claim. She suggests that people assent to a salient claim with a different content: not *that the environment will be harmed* but rather *that*

[4] Similar problems affect Schaffer's and Knobe's (2012) contrastive account of ESEE. They argue that, unlike in the morally positive cases, where the contrast tends to focus on epistemically relevant factors, in morally negative cases, the contrast is naturally taken to be between the morally negative action (the chairman's starting the program) and a morally positive or neutral action (not starting the program). Given that the story settles it that the chairman does start the program, the participants are justified in claiming that he knew that the environment would be harmed. However, it is not clear how to apply this explanation to the slight-chance of harm or the false-harm cases, where the chairman's action does not settle it that the environment would be harmed. Moreover, in the slight-chance of harm cases, it is not even clear that the chairman's starting the program is a morally negative action.

*a possible consequence of the chairman's action is that the environment will be harmed.* She distinguishes three knowledge claims: the knowledge claim people were asked about in the ESEE experiments (K), the predictive knowledge claim (P) and the consequence-knowledge claim (C):

(K) The chairman knew *that the environment would be harmed*.
(P) The chairman knew that [it is more likely than not] *that the environment would be harmed.*
(C) The chairman knew that a possible consequence of his action was *that the environment would be harmed*.

The content *that the environment would be harmed* is a part of the content of all three claims. It is a proper part of the content of (C) and (P). The remaining parts of the content of (C) and (P) make them independent of each another. One can know *for sure* what possible harmful consequence of one's action is without thereby being convinced at all that the consequence will probably occur. Likewise, a person might be convinced that it is likely that the environment will be harmed but she might not be aware that this will be a consequence of what she is about to do.

Moreover, the knowledge claim (K) could be used as an abbreviation of the other two claims. Arguably, (K) is naturally taken to be the predictive claim (P). The central thought of the consequence account of ESEE is that, in the harm cases, the knowledge claim (K) is interpreted as the consequence-awareness claim (C) rather than the predictive claim (P). Of course, the consequence-awareness claim is true both in the harm and in the help case. It can be argued, however, that it is more salient in the harm case, which would explain why people tend to attribute the knowledge claim in the harm case more often than in the help case.

There may be different explanations for the salience of the consequence-awareness claim in the harm case. Paprzycka-Hausman (2020) uses the omissions account (Paprzycka 2015, 2016) to argue that consequence awareness plays a pivotal role in attributing intentionality in the norm-violation cases. One could also invoke the sort of considerations raised by DHA to substantiate the salience of the awareness of consequences in norm-violation cases. Arguably, for practical reasons, it is rational to be epistemically alert in norm-violation situations, and so also to be more alert to those consequences of our behavior that may violate norms.

On the consequence account, the attribution of knowledge in norm-violation scenarios is affected by the claim about the awareness of consequences. Perhaps people could express themselves more precisely but there is clearly a rational component to their assertion. It is not clear to what extent precision of expression is a linguistic norm. Arguably, one should be as precise as to be understood in a given situation (cf. Grice 1989). When a surgeon yells "Scalpel!", the content communicated is usually settled by the situation. There is no need to demand a more precise expression.

The consequence account can explain attributions of knowledge in the slight-chance of harm and the false-harm cases. If the attribution of knowledge is (or is influenced by) the attribution of the consequence-awareness claim then knowledge should be attributed even in the case where the probability that the environment will be harmed is small, and indeed even if the environment is not ultimately harmed. In the slight-chance of harm cases, it is fully rational to claim that the chairman knew

what the possible consequence of his action was. For similar reasons, in Turri's (2014) study where the environment was not harmed, it is fully rational to claim that the chairman knew that a possible consequence of his action was that the environment would be harmed. This also shows that people's responses in the false-harm case need not be taken as evidence that people's concept of knowledge is not factive.

The consequence account thus solves the problems raised for the DHA explanation of ESEE. Moreover, the account has received some empirical support. It has been tested in the slight-chance of harm scenarios, in particular Butler-type stories (Paprzycka-Hausman 2020). In neutral stories, most participants were disinclined to attribute knowledge and cited the probability of the outcome (1/6) as their main justification for the denial. In the norm-violation stories, those participants who failed to attribute knowledge cited the probability of the outcome, but the majority of those who did attribute knowledge cited the fact that Brown was aware that a possible consequence of his action was that Smith would be killed (depending on the group, the majority ranged from 78 to 97%).

In sum, while the consequence account does not aspire to offer a unified account of all side-effect effects, it does provide a promising framework in which to explain ESEE. We conducted a study to adjudicate between the two accounts.

## 4 Study

On DHA, ESEE arises because it is rational to form beliefs about norm-violating side effects and thus people have a greater tendency to attribute beliefs in norm-violation cases than in norm-conformity cases. The question is what would happen if people had the requisite belief from sources other than the need to pay attention to norm violation. In other words, if the chairman had a sturdy initial belief that the environment would be helped rather than harmed in the harm case (e.g. because he is an optimist about the environment), *ceteris paribus* people ought to be disinclined to claim that the chairman knew that the environment would be harmed. By contrast, on the consequence account, people ought to attribute the knowledge claim in the harm condition even if the chairman had a "contrary" belief since the knowledge claim has a different (consequence-awareness) content.

### 4.1 Method

The study had a between-subject 2 (Outcome: the program harmed/helped the environment) × 2 (Belief: the chairman believes that the program will harm/help the environment) × 2 (Report: the chairman is told that the program would harm/help the environment) design. The vignettes were modifications of Knobe's harm and help scenarios (see Appendix 1). In each story, right after the vice-president announces, "We are thinking of starting a new program," an additional element concerning the chairman's psychology was added. The chairman was described either as an environmental pessimist who believes that any programs his company launches will harm the environment ($\text{Harm}^{\text{Bel}}$) or as an environmental optimist

**Table 1** A summary of the factors involved in the experimental conditions of the study

| Factor | Abbreviation | Description |
|---|---|---|
| Outcome | $Harm^{Out}$ | The program actually harmed the environment |
| | $Help^{Out}$ | The program actually helped the environment |
| Report | $Harm^{Rep}$ | According to the vice-president's report, the environment will be harmed |
| | $Help^{Rep}$ | According to the vice-president's report, the environment will be helped |
| Belief | $Harm^{Bel}$ | The chairman is an environmental pessimist and believes that any program his company launches will harm the environment |
| | $Help^{Bel}$ | The chairman is an environmental optimist and believes that any program his company launches will help the environment |

who believes that they help the environment ($Help^{Bel}$). This condition was meant to fix the chairman's belief as independent of what he is told by the vice-president who continues to say either that the program will harm ($Harm^{Rep}$) or help ($Help^{Rep}$) the environment. Since the chairman's belief is contradicted by what he is told by the vice-president in $Harm^{Rep}$-$Help^{Bel}$ and $Help^{Rep}$-$Harm^{Bel}$ conditions, it is added that the chairman was not swayed in his established beliefs but that he did not want to engage in a debate with the vice-president. In the remaining conditions, where the chairman's belief coincides with what he is told by the vice-president, it is explicitly stated that what the vice-president said confirms the chairman in his belief. Finally, the Outcome condition was indicated in the last sentence, i.e. whether the program harmed ($Harm^{Out}$) or helped ($Help^{Out}$) the environment. The following template for the vignettes was used (where the brackets mark the belief $Harm^{Bel}$/$Help^{Bel}$ condition, the braces mark the report $Harm^{Rep}$/$Help^{Rep}$ condition, and parentheses mark the outcome $Harm^{Out}$/$Help^{Out}$ condition, see also Table 1):

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program." (The chairman of the board is an environmental [*pessimist/optimist*]$^{Bel}$ – Taught by past experience, he believes that all programs his company will launch [*harm/help*]$^{Bel}$ the environment.)
> The vice-president said: "The program will help us increase profits, {*but/and*}$^{Rep}$ we have a report that it will also {*harm/help*}$^{Rep}$ the environment." [*The chairman was not persuaded. He still believed that the program will harm/help the environment but he did not want to engage in a debate/ This confirmed the chairman in his belief*]$^{Bel}$.
> The chairman answered, "I don't care at all about {*harming/helping*}$^{Rep}$ the environment. I just want to make as much profit as I can. Let's start the new program."
> They started the new program. As a result, the environment was actually (*harmed/helped*)$^{Out}$.

In one of the $Help^{Out}$ scenarios ($Help^{Out}$-$Harm^{Rep}$-$Harm^{Bel}$), the final sentence of the story was changed to account for the unexpected nature of the outcome ("As a result, quite unexpectedly, the environment was actually helped."). In

all the seven remaining groups, the stories ended without the addition of "quite unexpectedly". The vignette was visible on all screens.

On the first screen, subjects were asked to what extent they agree with the following claims:

(KHarm) The chairman knew that the program would harm the environment.
(KProfit) The chairman knew that the program would increase the company's profits.
(KHelp) The chairman knew that the program would help the environment.
(BHarm) The chairman believed that the program would harm the environment.
(BProfit) The chairman believed that the program would increase the company's profits.
(BHelp) The chairman believed that the program would help the environment.

The answers were recorded using a seven-point Likert scale ranging from -3 (labeled as "strongly disagree") to 3 (labeled as "strongly agree"). On consecutive screens, subjects were asked to justify their responses to (KHarm) and (KHelp). The following description pertains to the justification of (KHarm), see Appendix for the other.

Participants who agreed with the attribution of (KHarm), i.e. who marked a positive value, were presented with the following question and justification options (the letters in parentheses were not visible):

Which of the following statements best explains why you thought or tended to think that the chairman knew that the environment would be harmed:
(Q) The chairman was aware that it was possible that the new program would harm the environment.
(A) The chairman was aware that it was possible that the new program would help the environment.
(E) The chairman believed that there was sufficient evidence that the new program would harm the environment.
(B) The chairman believed that the new program would harm the environment.
(O) Other (Please explain).

The last option provided an opportunity to write in an answer. E-option was thought to be a standard justification for the attribution of knowledge. In both E-option and B-option, the chairman's propositional attitude was that of belief. Since both options were visible, it was assumed that participants who placed greater stress on evidence would choose E-option while those who placed greater stress on belief would choose B-option. It is noteworthy that the content of both E-option and B-option concerned the program's harming the environment in all conditions (including $\text{Help}^{\text{Bel}}$) conditions. This was because the knowledge claim in question (KHarm) concerned the program's harming the environment.

Participants who marked a negative value, i.e. disagreed with (KHarm), were asked an appropriately modified question with the above choices except that 'was aware' was replaced with 'wasn't aware' and 'believe' was replaced with 'didn't believe'. Finally, if they marked 0 (i.e. neither agreed nor disagreed with the claim), they were given the above choices except that the phrases 'The chairman [was aware/believed] that' were replaced with 'It wasn't clear whether the chairman [was or wasn't aware/believed or didn't believe] that' (see Table 2).

**Table 2** The contents of the justification options (Q, A, E, B, O) for (KHarm) attribution depending on participants' response to (KHarm)

| Option code | Option content | | | |
|---|---|---|---|---|
| | Responses to (KHarm) | | | |
| | +3, +2, +1 | 0 | -1, -2, -3 | |
| Q | The chairman was aware… | It wasn't clear whether the chairman was or wasn't aware… | The chairman wasn't aware… | …that it was possible that the new program would harm the environment |
| A | | | | …that it was possible that the new program would help the environment |
| E | The chairman believed… | It wasn't clear whether the chairman believed or didn't believe… | The chairman didn't believe… | …that there was sufficient evidence that the new program would harm the environment |
| B | | | | …that the new program would harm the environment |
| O | Other (please explain) | | | |

On the last screen, participants were asked whether they agreed or disagreed with several further claims (see Appendix 1). Responses were recorded using the same Likert scale. We use the answers to the following claims in our further discussion:

(AwHarm/AwHelp) The chairman was aware that it was possible that the new program would harm/help the environment.
(EvHarm/EvHelp) The chairman had sufficient evidence that the new program actually would harm/help the environment.
(JuHarm/JuHelp) The chairman thought that there was sufficient evidence that the new program actually would harm/help the environment.

Both the harm and the help versions of these claims were given in all eight groups.

### 4.2 Predictions

As far as the $\text{Harm}^{\text{Out}}$ scenarios were concerned, DHA would predict attributions of (KHarm) in the $\text{Harm}^{\text{Bel}}$ scenarios but not in the $\text{Help}^{\text{Bel}}$ scenarios. The key thought of DHA is that if people do not attribute the belief that the environment will be harmed, they will not attribute the knowledge that the environment will be harmed. To the extent that DHA accepts a standard account of knowledge, attributions of (KHarm) were not expected in the $\text{Help}^{\text{Out}}$ scenarios: the environment was not harmed after all.

The consequence account, on the other hand, would expect people to be mostly inclined to attribute (KHarm) in those $\text{Harm}^{\text{Out}}$ conditions, in which the chairman is aware of the possible consequence. The chairman is aware of it in $\text{Harm}^{\text{Rep}}$ groups (due to the vice-president's testimony) as well in $\text{Harm}^{\text{Bel}}$ groups (due to the chairman's own belief). In either case, the chairman is aware of the possible consequence that the environment will be harmed. In the $\text{Help}^{\text{Out}}$ scenarios, the fact that the environment was helped was expected to disincline people from the attribution of (KHarm). However, arguably, in the scenarios where the awareness of the consequence that the environment might be harmed is raised, one could expect somewhat heightened attributions of knowledge. In other words, in the $\text{Help}^{\text{Out}}$-$\text{Harm}^{\text{Rep}}$ conditions as well as in the $\text{Help}^{\text{Out}}$-$\text{Harm}^{\text{Bel}}$ conditions, one could expect the ultimate judgment that people give to be affected by two sources: people's disinclination to attribute knowledge that the environment would be harmed (because the environment was ultimately helped) and people's inclination to attribute knowledge of consequences.

**Table 3** Predictions of DHA and the consequence account with respect to the attribution of (KHarm)

| | | DHA | | Consequence | |
|---|---|---|---|---|---|
| | | $\text{Harm}^{\text{Bel}}$ | $\text{Help}^{\text{Bel}}$ | $\text{Harm}^{\text{Bel}}$ | $\text{Help}^{\text{Bel}}$ |
| $\text{Harm}^{\text{Out}}$ | $\text{Harm}^{\text{Rep}}$ | + | – | + | + |
| | $\text{Help}^{\text{Rep}}$ | + | – | + | – |
| $\text{Help}^{\text{Out}}$ | $\text{Harm}^{\text{Rep}}$ | – | – | +/– | +/– |
| | $\text{Help}^{\text{Rep}}$ | – | – | +/– | – |

The predictions of both accounts coincide to a large extent (see Table 3). They were expected to differ distinctly in one group ($Harm^{Out}$-$Harm^{Rep}$-$Help^{Bel}$) and to some extent in three other groups ($Help^{Out}$-$Harm^{Rep}$-$Harm^{Bel}$, $Help^{Out}$-$Harm^{Rep}$-$Help^{Bel}$, $Help^{Out}$-$Help^{Rep}$-$Harm^{Bel}$). The most pronounced difference concerns the prediction in the $Harm^{Out}$-$Harm^{Rep}$-$Help^{Bel}$ condition. According to DHA's central tenet (no belief, so no knowledge), the chairman's belief that the environment will be helped (i.e. not harmed) ought to disincline people from (KHarm) attribution. According to the consequence account, despite the fact that the chairman believes that the environment will be helped, he is aware that it might be harmed. Since, on that account, the consequence-awareness claim is largely responsible for the attribution of (KHarm), one should expect people to attribute it in this group.

In the $Help^{Out}$ groups, to the extent that DHA and the consequence account accept a standard conception of knowledge, they would presumably not predict the attribution of (KHarm) since the truth condition is not satisfied (the environment is helped rather than harmed). However, the consequence account sees another source for the attribution of (KHarm), viz. consequence awareness, which is present in the $Harm^{Rep}$ conditions and in the $Harm^{Bel}$ conditions. It could be expected to influence the knowledge attribution somewhat.

On the consequence account, it was also expected that people would tend to justify the attribution of (KHarm) by appeal to the awareness of consequences (Q-option) rather than to the evidence (E-option) or to the chairman's beliefs (B-option). This was to be expected in the conditions where the possibility of harming the environment was salient, especially $Harm^{Rep}$ conditions. In the $Harm^{Bel}$ conditions, where the possibility of harm was salient due to the chairman's beliefs, the knowledge attribution could also be justified by appeal to the chairman beliefs (especially in the $Help^{Out}$-$Help^{Rep}$ condition). The consequence account does not make any clear predictions concerning the justifications of the disagreement with (KHarm). To the extent that it relies on a standard account of knowledge, it would expect that people would justify their disagreement by appealing to the fact that one of the standard conditions on knowledge was not satisfied: belief (B-option), justification (E-option), or truth (which could be entered in the open option).

On DHA, on the other hand, one would expect that people will justify their disagreement with (KHarm) by appealing to the fact that the chairman lacked the belief (B-option). It is less clear on DHA how respondents should justify their attributions of knowledge. Insofar as DHA takes belief to be central to the attributions of knowledge, one might expect participants to justify such attributions by appeal to the belief that the environment would be harmed (B-option). Insofar as DHA can explain the attributions of knowledge by appealing to the satisfaction of the classical concept of knowledge, one could perhaps argue that participants might choose also other options such as the belief about sufficient evidence option (E-option) or possibly truth. However, on such an interpretation, one would still expect B-option to be the most frequent justification for DHA. If the doxastic heuristic is indeed employed, the belief condition will be the most salient of all three knowledge conditions.

### 4.3 Participants

429 participants (276 females; age: M = 35, SD = 12.5) took part in the study. The study was conducted using Prolific. Subjects were financially compensated (£ 0.8 for 7-min survey). Three people failed a simple attention check at the beginning of the study. Their exclusion did not affect the results, so we decided to include them in the final sample.

### 4.4 Results

**Knowledge Attributions** Analysis of variance of the (KHarm) responses (Fig. 1) revealed a significant effect of Outcome ($Harm^{Out}$ vs. $Help^{Out}$: $F(1,421) = 43.95$, $p < 0.001$, $\eta^2_p = 0.12$), Belief ($Harm^{Bel}$ vs. $Help^{Bel}$: $F(1,421) = 88.20$, $p < 0.001$, $\eta^2_p = 0.19$) and Report ($Harm^{Rep}$ vs. $Help^{Rep}$: $F(1,421) = 101.69$, $p < 0.001$, $\eta^2_p = 0.21$). No interactions were statistically significant.[5]

For (KHelp), the results were similar (Fig. 1). We found a statistically significant effect of Outcome ($F(1,421) = 20.41$, $p < 0.001$, $\eta^2_p = 0.07$), Belief ($F(1,421) = 121.06$, $p < 0.001$, $\eta^2_p = 0.22$) and Report ($F(1,421) = 83.08$, $p < 0.001$, $\eta^2_p = 0.20$). We also found a small but statistically significant interaction between Outcome and Report ($F(1,421) = 4.35$, $p = 0.038$, $\eta^2_p = 0.01$) as well as Outcome and Belief ($F(1,421) = 5.37$, $p = 0.021$, $\eta^2_p = 0.01$).[6]

Figure 2 compares the attributions of (KHarm) in the $Harm^{Out}$ condition with the attributions of (KHelp) in the $Help^{Out}$ condition. The original ESEE (Beebe and Buckwalter 2010) is thus replicated despite the more complex set-up.[7] There is a greater tendency to attribute knowledge in the $Harm^{Out}$ (M = 0.99, SD = 2.05) than in the $Help^{Out}$ (M = -0.58, SD = 2.18) conditions ($F(1,421) = 98.00$, $p < 0.001$, $\eta^2_p = 0.19$).[8]

DHA predicts that (KHarm) will not be attributed in $Help^{Bel}$ cases where the chairman does not believe that the environment will be harmed.[9] This includes the

[5] Outcome×Report: $F(1,421) = 0.0003$, $p = 0.978$; Outcome×Belief: $F(1,421) = 0.118$, $p = 0.731$; Report×Belief: $F(1,421) = 0.025$, $p = 0.875$; Outcome×Report×Belief: $F(1,421) = 2.39$, $p = 0.113$.

[6] Report×Belief: $F(1,421) = 0.700$, $p = 0.403$; Outcome×Report×Belief: $F(1,421) = 0.323$, $p = 0.570$.

[7] Since our set-up differed substantially from Knobe's original scenarios, we have also conducted two additional analyses. First, we compared only those conditions where the actual outcome and vice-president's report were congruent (4 groups). For those groups, the difference between $Harm^{Out}$ (M = 1.83, SD = 1.60) and $Help^{Out}$ (M = 0.31, SD = 2.10) was even more pronounced. Analysis of variance revealed a significant effect of Outcome ($F(1,225) = 53.18$, $p < 0.001$), Belief ($F(1,225) = 5.27$, $p = 0.023$) and their interaction ($F(1,225) = 69.58$, $p < 0.001$). Second, we compared 2 groups that most closely resembled the original Beebe and Buckwalter (2010) experiment ($Harm^{Out}$-$Harm^{Rep}$-$Harm^{Bel}$ as the Harm condition and $Help^{Out}$-$Help^{Rep}$-$Help^{Bel}$ as the Help condition). Our analysis showed again that the difference in knowledge attributions was statistically significant (Harm: M = 2.50 SD = 0.97, Help: M = 1.19, SD = 1.82, $t(128) = 5$, $p < 0.001$).

[8] We also found statistically significant interactions between Outcome and Report ($F(1,421) = 106.56$, $p < 0.001$, $\eta^2_p = 0.20$) as well as Outcome and Belief ($F(1,421) = 133.00$, $p < 0.001$, $\eta^2_p = 0.24$).

[9] One might perhaps worry that the chairman might have two beliefs: that the environment will be helped and that the environment will be harmed. This proved not to be the case. Both beliefs were attributed to the chairman only by 18 out of 421 respondents (by 2 out of 52 respondents in the crucial $Harm^{Out}$-$Harm^{Rep}$-$Help^{Bel}$ group).

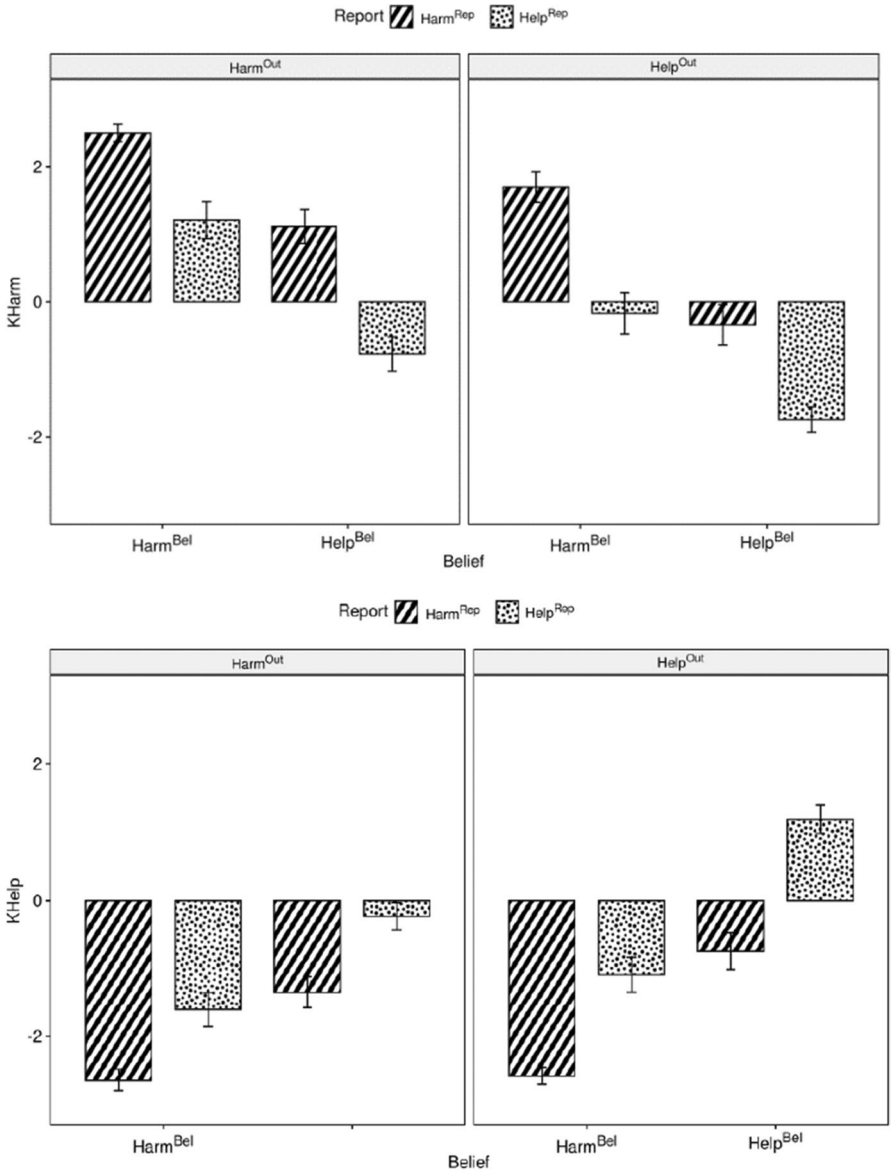


**Fig. 1** Attributions of knowledge of the negative (KHarm) and the positive (KHelp) side-effect (error bars represent standard error of mean)

$Harm^{Out}$-$Harm^{Rep}$-$Help^{Bel}$ group. By contrast, the consequence account predicts that (KHarm) will be attributed in this group due to the salience of the consequence-awareness claim. As Table 4 shows, participants in fact do tend to attribute (KHarm) in this case ($M=1.115$, $SD=1.822$, $t(51)=4.441$, $d=0.612$), which supports the consequence account.

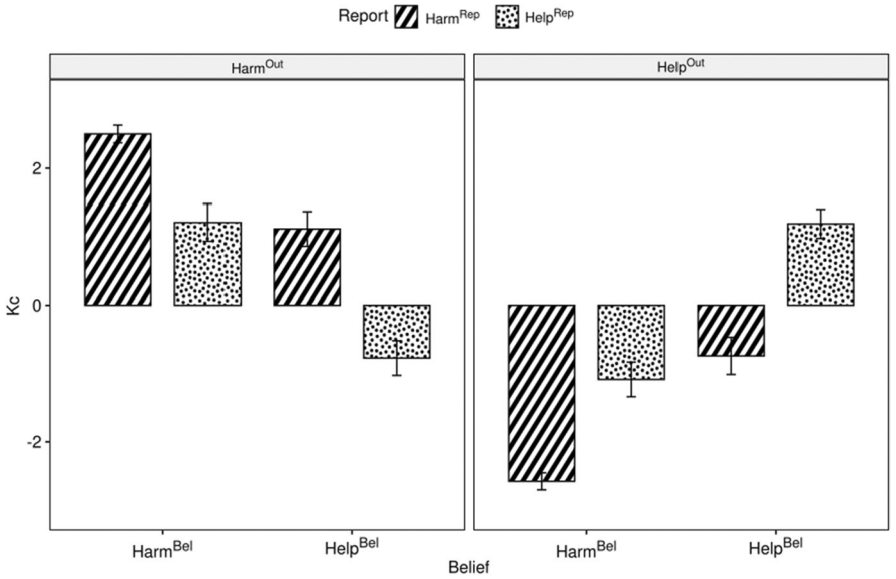


**Fig. 2** Attributions of (KHarm) in Harm$^{Out}$ conditions and (KHelp) in Help$^{Out}$ conditions; error bars represent standard error of mean

As we have seen, the predictions of the two accounts also diverge in cases where the truth condition of knowledge is not satisfied but the consequence-awareness claim may still be salient due to the report or to the agent's belief (Help$^{Out}$-Harm$^{Rep}$-Harm$^{Bel}$, Help$^{Out}$-Harm$^{Rep}$-Help$^{Bel}$, and Help$^{Out}$-Help$^{Rep}$-Harm$^{Bel}$). Insofar as both DHA and the consequence account accept a classical concept of knowledge, (KHarm) should not be attributed in view of the fact that the environment is helped (the truth condition is not satisfied). However, according to the consequence account, the salience of the consequence-awareness claim (due to what is reported or to what the chairman believes) may heighten the attributions of (KHarm). In fact (Table 4), mean knowledge attributions did not differ significantly from the midpoint (4) in Help$^{Out}$-Harm$^{Rep}$-Harm$^{Bel}$ and in Help$^{Out}$-Help$^{Rep}$-Harm$^{Bel}$ (though they are significantly higher than in Help$^{Out}$-Help$^{Rep}$-Help$^{Bel}$) but we observed a strong tendency in the direction of (KHarm) attribution ($M = 1.7$, $SD = 1.594$, $t(49) = 7.541$, $d = 1.067$) in Help$^{Out}$-Harm$^{Rep}$-Harm$^{Bel}$. In sum, the results from these three conditions do not support DHA but they are consistent with the predictions of the consequence account.

**Relation between knowledge and belief attributions** DHA explains the asymmetry in knowledge attributions in terms of the asymmetry in belief attributions. Since belief is a necessary condition for knowledge, people should attribute (or refuse to attribute) belief more strongly than they attribute (or refuse to attribute) knowledge. By contrast, the consequence account allows for the possibility of

**Table 4** Descriptive statistics for (KHarm) attribution (for each condition one-sample *t*-test is reported; $H_0$: μ = 0)

| Outcome | Report | Belief | *n* | M | SD | *t* | *df* | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|---|---|---|
| Harm | Harm | Harm | 56 | 2.5 | 0.972 | 19.24 | 55 | < 0.001 | 2.571 |
| | | Help | 52 | 1.115 | 1.822 | 4.414 | 51 | < 0.001 | 0.612 |
| | Help | Harm | 43 | 1.209 | 1.794 | 4.421 | 42 | < 0.001 | 0.674 |
| | | Help | 57 | -0.772 | 1.918 | -3.038 | 56 | 0.004 | -0.402 |
| Help | Harm | Harm | 50 | 1.7 | 1.594 | 7.541 | 49 | < 0.001 | 1.067 |
| | | Help | 50 | -0.34 | 2.125 | -1.131 | 49 | 0.263 | -0.16 |
| | Help | Harm | 47 | -0.17 | 2.099 | -0.556 | 46 | 0.581 | -0.081 |
| | | Help | 74 | -1.743 | 1.631 | -9.196 | 73 | < 0.001 | -1.069 |

knowledge attributions without belief attributions. This will be the case when the consequence-awareness claim is salient. The key condition in this regard is the $\text{Harm}^{\text{Out}}$-$\text{Harm}^{\text{Rep}}$-$\text{Help}^{\text{Bel}}$ condition since the chairman does not have the belief (that the new program will harm the environment) but the consequence-awareness claim is salient due the vice-president's testimony (the chairman knows that a possible consequence of his action is that the program will harm the environment).

In fact (Table 5), there is a significantly stronger tendency to attribute knowledge (M = 1.115) than belief (M = -0.115; $t(51) = 4.56$, $p < 0.001$, $d = -0.632$) in this scenario. This result indicates that there is a considerable number of participants who attributed knowledge more firmly than belief. It thus suggests that knowledge attributions do not depend entirely on belief attributions. These results are inconsistent with the core thought of DHA.

To investigate the relation between knowledge and belief attribution, we computed Pearson's correlation coefficients for knowledge attribution (KHarm/KHelp) as one variable and belief attribution (BHarm/BHelp), justification attribution (JuHarm/JuHelp), consequence-awareness attribution (AwHarm/AwHelp) and evidence attribution (EvHarm/EvHelp) as a second variable. If the DHA

**Table 5** Differences between (KHarm) and (BHarm) attributions (means are reported; for each condition paired *t*-test is reported)

| Outcome | Report | Belief | *n* | KHarm | BHarm | *t* | *df* | *p* | Cohen's *d* |
|---|---|---|---|---|---|---|---|---|---|
| Harm | Harm | Harm | 56 | 2.5 | 2.696 | 1.375 | 55 | 0.175 | 0.184 |
| | | Help | 52 | 1.115 | -0.1154 | -4.56 | 51 | < 0.001 | -0.632 |
| | Help | Harm | 43 | 1.209 | 1.791 | 1.871 | 42 | 0.068 | 0.285 |
| | | Help | 57 | -0.772 | -1.123 | -1.507 | 56 | 0.137 | -0.200 |
| Help | Harm | Harm | 50 | 1.7 | 2.34 | 2.644 | 49 | 0.011 | 0.374 |
| | | Help | 50 | -0.34 | -0.16 | 0.837 | 49 | 0.407 | 0.118 |
| | Help | Harm | 47 | -0.17 | 1.894 | 5.747 | 46 | < 0.001 | 0.838 |
| | | Help | 74 | -1.743 | -2.027 | -2.186 | 73 | 0.032 | -0.254 |

explanation of ESEE is correct, we should expect the relation between belief and knowledge attribution to be the strongest of all mentioned above. On the other hand, the consequence account predicts that the attribution of the consequence-awareness claim should be at least as important as the attribution of belief, in particular for (KHarm).

In fact, in the $\text{Harm}^{\text{Out}}$ conditions, consequence awareness exhibits a stronger correlation with knowledge than does belief ($r=0.74$ vs. $r=0.64$, see also Table 7 in Appendix 2). Although this difference is relatively small, it favors the consequence account over DHA. Using Steiger's (1980) approach, we compared the two correlation coefficients and found them to be significantly different ($z=2.594$, $p=0.01$). In the $\text{Help}^{\text{Out}}$ conditions, there were no statistically significant differences between the correlations of knowledge attributions with belief and consequence-awareness attributions.

**Justifications** Recall that participants could choose four claims to justify their response or they could offer another justification. For all positive, negative and hesitant responses (see Table 8 in Appendix 2), the consequence-awareness Q-option was chosen by 31%, the other awareness A-option was chosen by 10%, the belief about sufficient evidence E-option was chosen by 17%, the belief B-option was chosen by 33%, and another justification was offered by 8% of participants.

According to the consequence account, participants should justify their attribution of (KHarm) by appeal to Q-option. Participants who failed to attribute knowledge were expected to appeal to the fact that some of the standard conditions on knowledge were not satisfied. DHA, on the other hand, predicts that people would justify their disagreement with knowledge attribution by appealing to the lack of belief (B-option). As we argued above, B-option should also be the most salient for those who attribute knowledge, even though the other conditions on knowledge may also be relevant.

We trichotomized (KHarm) responses into three categories: "yes" ($>0$), "no" ($<0$) and "neither" (0). We then compared the distribution of justifications indicated by participants in each category. The results are presented in Fig. 3 (see also Table 8 in Appendix 2). We found that almost 50% of respondents who attributed knowledge appealed to consequence awareness as a justification (it was exactly 50% in the $\text{Harm}^{\text{Out}}$ conditions). Among the participants who did not attribute knowledge, the most frequent justification (42%) referred to the agent's lack of the relevant belief. This result suggests that the adequacy of the consequence-awareness claim is an important factor that "pushes" the participants to attribute (KHarm), which is in line with the consequence account. It should be stressed, however, that the results for the "no" category align with the DHA account. The most frequently chosen justification for the failure to attribute knowledge was the lack of belief.

In order to investigate the relationship between knowledge attributions and justifications even further, we compared two linear regression models (see Table 6). In both models, the attribution of knowledge is the predicted variable. In the first model ("DHA-model"), we entered only one predictor – belief attribution (BHarm), which should

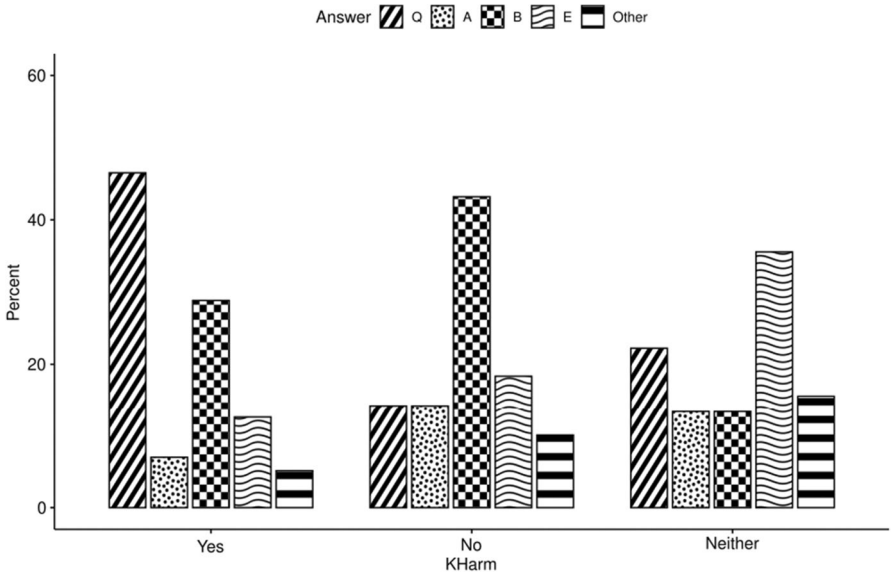


**Fig. 3** Justifications of participants' attitude to (KHarm)

explain ESEE according to DHA. In the second model ("CA-model"), we also included justifications of (KHarm) as a categorical variable and interaction terms.[10] First, in the CA model, we found a statistically significant effect of (KHarm) justification, which may indicate that some part of knowledge attribution is not explained by belief attribution. Participants who appealed to the consequence-awareness claim on average exhibited stronger knowledge attribution than those who appealed to belief ($b=2.09$, $p<0.001$). Second, in the CA model, interaction terms appear to be statistically significant (see Fig. 4). Negative regression coefficients (BHarm×E-Option: $b=-0.38$, $p<0.001$; BHarm×Q-Option: $b=-0.29$, $p<0.001$) suggest that the belief of the agent was less important for a significant portion of our sample who appealed to these two justification options. Finally, the CA-model fits the data significantly better than the DHA-model ($\Delta R^2=0.069$, $p<0.001$).

### 4.5 Discussion

In the study, we have shown that the DHA explanation of ESEE in terms of the chairman's belief is not sufficient. Two findings support this conclusion. First, in some conditions, people attribute knowledge of the side effect even in cases where the chairman is said not to have the relevant belief (but a contrary belief). The fact that participants are willing to ascribe knowledge more strongly than belief contradicts the core thought of DHA. Second, we compared two models

[10] In coding the categorical variable, we used B-option as a base category. We also excluded all participants who appealed to the A-option (too few observations) or selected the "Other"-option (non-homogenous group).

**Table 6** Regression results using (KHarm) as the criterion

| Predictor | *b* | *b* 95% CI [LL, UL] | $sr^2$ | $sr^2$ 95% CI [LL, UL] | Fit — Difference |
|---|---|---|---|---|---|
| (Intercept) | 1.18** | [0.85, 1.52] | | | $R^2 = 0.582$** |
| BHarm | 0.74** | [0.67, 0.80] | 0.58 | [0.52, 0.63] | 95% CI[0.52, 0.63] |
| (Intercept) | 0.37 | [-0.04, 0.79] | | | |
| BHarm | 0.88** | [0.79, 0.97] | 0.39 | [0.32, 0.47] | $R^2 = 0.651$** |
| E-Option | 1.38** | [0.55, 2.21] | 0.01 | [-0.00, 0.02] | 95% CI[0.59, 0.69] |
| Q-Option | 2.09** | [1.37, 2.81] | 0.03 | [0.01, 0.06] | — |
| BHarm×E-Option | -0.38** | [-0.55, -0.22] | 0.02 | [0.00, 0.04] | $\Delta R^2 = 0.069$** |
| BHarm×Q-Option | -0.29** | [-0.43, -0.15] | 0.02 | [0.00, 0.03] | 95% CI[0.04, 0.10] |

A significant *b* indicates that the semi-partial correlation is also significant. *b* represents unstandardized regression weights. $sr^2$ represents the semi-partial correlation squared. *LL* and *UL* indicate the lower and upper limits of the confidence interval, respectively. * indicates $p < 0.05$. ** indicates $p < 0.01$.

of knowledge attribution: the DHA-model (*belief* predicts *knowledge*) and the CA-model (*belief* combined with *justification* predicts *knowledge*). This comparison indicated that the addition of different justification options significantly contributed to the predictive power of the statistical model. Taken together with the theoretical objections we raised (§2), the results suggest that DHA is unable to account for important aspects of knowledge attribution in ESEE cases.

Our findings also provide some support for the consequence account. The consequence account can explain attributions of knowledge not only where the truth (Turri 2014) and the justification (Turri 2014; Beebe and Jensen 2012; Paprzycka-Hausman 2020) conditions are not satisfied, but also where belief is absent. One can disagree that the chairman believed that the environment would be harmed and yet agree that the chairman knew that a possible consequence of starting the program was that the environment would be harmed.

Another piece of evidence that favours the consequence account is the pattern of responses to the justification question. In the study (especially in the $\text{Harm}^{\text{Out}}$ condition), Q-option was the most frequently chosen justification by participants who attributed knowledge to the chairman, while B-option was the predominant justification for those participants who did not attribute knowledge. The consequence account can explain this pattern very cleanly. Due to the salience of norm violation, some participants tracked the chairman's awareness of the consequences of his actions and thus tended to attribute knowledge. Other respondents who interpreted the knowledge claim literally tracked the chairman's beliefs (with a predictive content, i.e. beliefs that the side effect will occur), their truth and justification. Since the truth condition was satisfied in the $\text{Harm}^{\text{Out}}$ condition and the justification condition was arguably also satisfied, they denied knowledge because the chairman did not have the belief.

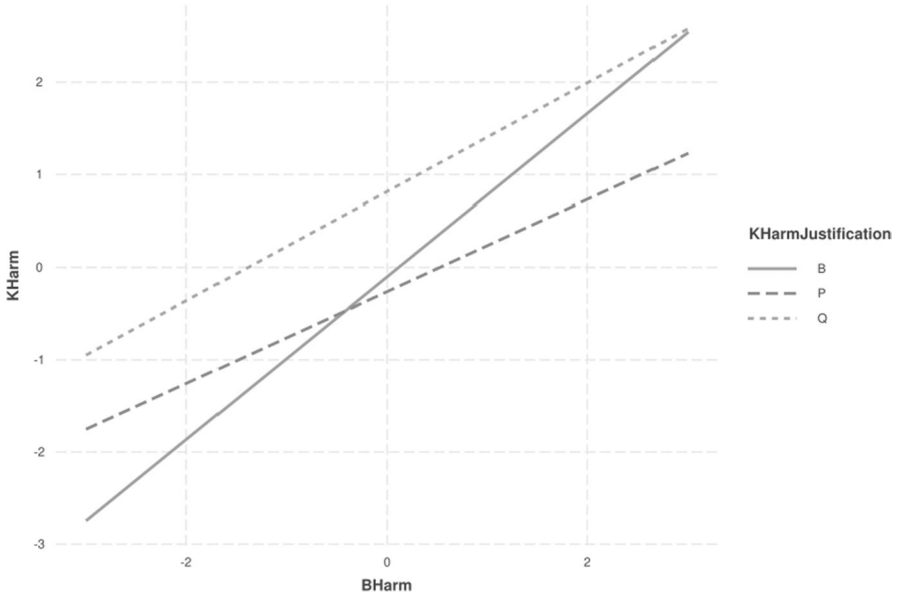


**Fig. 4** Relation between belief and knowledge attribution for three justification options: belief (B), belief about sufficient evidence (E), and consequence awareness (Q)

By contrast, DHA seems to be unable to account for this pattern. DHA can, of course, explain why the lack of belief was the most frequently chosen justification by participants who failed to attribute knowledge (44% of participants chose B-option in the $\text{Help}^{\text{Out}}$ conditions). However, it is arguable that DHA would also expect B-option to be the most frequently chosen in the $\text{Harm}^{\text{Out}}$ conditions. After all, DHA takes belief to be central to the attributions of knowledge. Moreover, even if the other conditions (justification, truth) are relevant, the belief option should be the most salient if the doxastic heuristic is indeed employed. We see, however, that this is not the case: 27% of participants chose B-option in the $\text{Harm}^{\text{Out}}$ conditions.

One limitation of our study is that it does not provide direct evidence for the consequence account. Rather the consequence account can be thought of as the best explanation of the pattern of responses emerging from the data. This is a serious concern that would require further experimental investigation, possibly also the employment of a more innovative methodological approach that goes beyond the questionnaire method.

Second, a possible objection is that many of our claims hinge on the ability of the participants to reflect on their knowledge attributions. Such reflection obviously requires a non-trivial amount of epistemological competence, which the subjects may lack. DHA proponents might argue that the justifications provided by the subjects are thus unreliable. However, our justification question should not be understood as a question about the alleged psychological causes of knowledge attribution. Arguably, people do not have the capacity for this kind of self-reflection. The answers to questions about justification are frequently

post hoc reconstructions. They are questions about reasons not causes. As such, the subjects' answers can be taken to be revealing about the way they think of knowledge in the scenarios presented to them. They are the data that a philosophical account of ESEE should report on.

Third, one may raise objections related to study design.[11] Given the importance of the attributions of knowledge and belief, the questions on Screen 1 could have been randomized. We acknowledge that this is a potential problem. We have rerun the most critical group of the study ($Harm^{Out}$-$Harm^{Rep}$-$Help^{Bel}$) where the order of questions on Screen 1 was changed – belief questions were asked before knowledge questions. We still obtained the most striking result of the study – there was a statistically significant difference between the attribution of KHarm (M = 0.612, SD = 1.86) and BHarm (M = -0.163, SD = 2.11; paired *t*-test: $t(48) = 3.15$, $p = 0.003$). There were no major differences between the original group and the group where the question order was changed,[12] which might indicate that the order was not critical to the study after all.

Another problem concerns the way in which the long standing belief is introduced in the story. The chairman's belief that the environment will be helped (harmed) is introduced in the context of his being described as an optimist (pessimist). One might worry that attitudes such as optimism are associated with hopes or fears rather than beliefs. As a result, participants might attribute a relatively weak notion of belief to the chairman. We reran the most critical group ($Harm^{Out}$-$Harm^{Rep}$-$Help^{Bel}$) with a modification of the story, in which the reference to optimism was dropped.[13] The fragment "The chairman of the board is an environmental optimist. – Taught by past experience, he believes that all programs his company launches help the environment" was replaced by "Taught by long experience, the chairman of the board believes that all programs his company launches help the environment." The striking difference in the attribution of KHarm (M = 1.64, SD = 1.42) and BHarm (M = 0.26, SD = 1.89) was also statistically significant in the rerun study (paired *t*-test: $t(46) = 5.53$, $p < 0.001$). There were no other evident differences between the original group and the group with a modified story.[14]

[11] We would like to thank one of the anonymous reviewers for raising these objections.

[12] Although it is generally not recommended to run statistical tests on the data from different studies, we decided to compare the results of both reruns to the original study as a robustness check. The differences between the original condition and the rerun with reversed question order ($n = 49$) were not significant for both KHarm ($t(98) = 1.90$, $p = 0.06$) and BHarm ($t(99) = 0.90$, $p = 0.37$).

[13] There is another reason for dropping the reference to optimism. As one of our reviewer's pointed out, the description of the chairman as an environmental optimist stands in some tension with the later declaration of indifference about the environment. In our two reruns, we have asked participants to what extent they agree that the chairman did not care about the environment. (This question was asked on the last screen together with many other claims but was not asked in the original study.) In the rerun group with the modified story ("optimism" dropped), people were inclined to agree that the chairman didn't care (M = 1.21, SD = 1.78, $n = 47$), in the group with the original story, people were somewhat inclined to agree that the chairman didn't care (M = 0.45, SD = 2.13, $n = 49$), though the difference between the groups was not statistically significant ($t(92) = 1.91$ $p = 0.06$).

[14] We performed a robustness check for the attributions of KHarm and BHarm (see also footnote 12). The differences between the original condition and the rerun with "optimism" dropped ($n = 47$) were not significant (for KHarm: $t(97) = 0.89$, $p = 0.37$; for BHarm: $t(97) = 0.39$, $p = 0.700$).

It should be acknowledged that the story is rather complex and that there are accordingly potential problems. However, the results obtained in the two reruns of the $Harm^{Out}$-$Harm^{Rep}$-$Help^{Bel}$ condition proved rather stable. In all the reruns, participants were inclined to attribute the knowledge that the environment would be harmed even though they were disinclined to attribute the belief that the environment would be harmed. This result is inconsistent with the core thought of DHA.

It should be acknowledged that the consequence account lacks the generality of DHA. The latter can (at least potentially) explain a wide range of asymmetries in the attribution of different kinds of mental states; the former focuses on asymmetries in the attribution of knowledge. Generality is without a doubt an important theoretical virtue (cf. Hindriks 2019). However, we have shown that DHA has problems in explaining some experimental data including the data presented in this paper.

## 5 Conclusion

We have considered two accounts of ESEE: DHA and the consequence account. While DHA promises to be a more comprehensive account, it has problems with accounting for some of the existing data. In particular, it has problems with explaining knowledge attributions in cases where the probability that the side effect will occur is low and in cases where the side effect does not ultimately occur. DHA can explain why there is a difference between the harm and the help cases but it cannot explain why people are willing to attribute knowledge in the harm cases when the justification or the truth condition is not satisfied.

The consequence account is limited to the explanation of ESEE but it can explain the attributions of knowledge in the problematic harm cases. Its central thesis is that such knowledge attributions are due to the increased salience of the consequence-awareness claim. In being prepared to claim that the chairman knew that the environment would be harmed, people give expression to their view that he was aware that the possible consequence of his action was that the environment would be harmed. Such a claim can be true even if the probability that the consequence will occur is low and even if the consequence does not occur. According to the consequence account, attributions of knowledge in harm cases (even if they are not matched by attributions of knowledge in help cases) are at root rational though people attribute a different knowledge claim in the norm-violation cases.

Our results show that people tend to justify their attributions of knowledge in the harm scenarios predominantly by appeal to the chairman's awareness of consequences. The predictions of DHA and of the consequence account were confronted. The results fit the predictions of the consequence account both in the overall pattern of responses as well as in the appeal to consequence awareness as a justification for knowledge attribution. However, the appeal to belief was also a significant factor in the study. It may well be that a complete explanation of ESEE should include both factors.

## Appendix 1 Study Design

The study had a between-subject 2 ($Harm^{Out}$ vs. $Help^{Out}$)$\times$2 [$Harm^{Bel}$ vs. $Help^{Bel}$]$\times$2 {$Harm^{Rep}$ vs. $Help^{Rep}$} design.

**Screen 1**

Respondents were presented with the following question and answer options (the letters in parentheses were not visible):

[Story]
To what extent do you agree with the following claims:
(KHarm) The chairman knew that the program would harm the environment.
(KProfit) The chairman knew that the program would increase the company's profits.
(KHelp) The chairman knew that the program would help the environment.
(BHarm) The chairman believed that the program would harm the environment.
(BProfit) The chairman believed that the program would increase the company's profits.
(BHelp) The chairman believed that the program would help the environment.

Participants specified their level of agreement regarding the chairman's knowledge / beliefs using a horizontal seven-point Likert scale ranging from -3 to + 3 (-3 was labeled as 'strongly disagree' and + 3 as 'strongly agree').

**Screen 2**

**[Story]**

**Did the chairman intentionally harm the environment?**

Yes/No

**How much blame does the chairman deserve for harming the environment?**

Participants specified their level of blame attribution using a horizontal seven-point Likert scale ranging from 1 to 7 (1 was labeled as 'No blame' and 7 as 'Very much blame').

**Screen 3**

[Story]

Participants were presented with the following questions and justification options (the letters in parentheses were not visible) – depending on whether they agreed, disagreed, or neither agreed nor disagreed with (KHarm):

a) Agreement (answers ranging from + 3 to + 1) with the attribution of (KHarm)

Which of the following statements best explains why you agreed (or tended to agree) that **the chairman knew that the environment would be harmed**:

(Q) The chairman was aware that it was possible that the new program would harm the environment.
(A) The chairman was aware that it was possible that the new program would help the environment.
(E) The chairman believed that there was sufficient evidence that the new program would harm the environment.
(B) The chairman believed that the new program would harm the environment.
(O) Other (Please explain).

b) Disagreement (answers ranging from -3 to -1) with the attribution of (KHarm)

Which of the following statements best explains why you disagreed (or tended to disagree) that **the chairman** knew **that the environment would be harmed**:

(Q) The chairman wasn't aware that it was possible that the new program would harm the environment.
(A) The chairman wasn't aware that it was possible that the new program would help the environment.
(E) The chairman didn't believe that there was sufficient evidence that the new program would harm the environment.
(B) The chairman didn't believe that the new program would harm the environment.
(O) Other (Please explain).

iii) Neither agreement nor disagreement (answers labeled as 0) with the attribution of (KHarm)

Which of the following statements best explains why you neither agreed nor disagreed with the claim that **the chairman knew that the new program would harm the environment**:

(Q) It wasn't clear whether the chairman was or wasn't aware that it was possible that the new program would harm the environment.
(A) It wasn't clear whether the chairman was or wasn't aware that it was possible that the new program would help the environment.
(E) It wasn't clear whether the chairman believed or didn't believe that there was sufficient evidence that the new program would harm the environment.
(B) It wasn't clear whether the chairman believed or didn't believe that the new program would harm the environment.
(O) Other (Please explain).

Respondents were asked to choose only one justification option.

**Screen 4**

Participants were presented with the following questions and justification options (the letters in parentheses were not visible) – depending on their response to (KHelp).

a) Agreement (answers ranging from + 3 to + 1) with the attribution of (KHelp)

Which of the following statements best explains why you agreed (or tended to agree) that **the chairman knew that the environment would be helped**:

(Q) The chairman was aware that it was possible that the new program would help the environment.
(A) The chairman was aware that it was possible that the new program would harm the environment.
(E) The chairman believed that there was sufficient evidence that the new program would help the environment.
(B) The chairman believed that the new program would help the environment.
(O) Other (Please explain).

b) Disagreement (answers ranging from -3 to -1) with the attribution of (KHelp)

Which of the following statements best explains why you disagreed (or tended to disagree) that **the chairman** knew **that the environment would be helped**:

(Q) The chairman wasn't aware that it was possible that the new program would help the environment.
(A) The chairman wasn't aware that it was possible that the new program would harm the environment.
(E) The chairman didn't believe that there was sufficient evidence that the new program would help the environment.
(B) The chairman didn't believe that the new program would help the environment.
(O) Other (Please explain).

iii) Neither agreement nor disagreement (answers labeled as 0) with the attribution of (KHelp)

Which of the following statements best explains why you neither agreed nor disagreed with the claim that **the chairman knew that the new program would help the environment**:

(Q) It wasn't clear whether the chairman was or wasn't aware that it was possible that the new program would help the environment.
(A) It wasn't clear whether the chairman was or wasn't aware that it was possible that the new program would harm the environment.
(E) It wasn't clear whether the chairman believed or didn't believe that there was sufficient evidence that the new program would help the environment.
(B) It wasn't clear whether the chairman believed or didn't believe that the new program would help the environment.
(O) Other (Please explain).

Respondents were asked to choose only one justification option.

**Screen 5**

Participants were presented with the following question and answer options (the letters in parentheses were not visible):

[Story]
To what extent do you agree with the following claims:
(Aw-) The chairman was aware that it was possible that the new program would…
(-Harm) … harm the environment,
(-Profit) … increase profits,
(-Help) … help the environment.
(Ev-) The chairman had sufficient evidence that the new program actually would…
(-Harm) … harm the environment,
(-Profit) … increase profits,
(-Help) … help the environment.

(Ju-) The chairman thought that there was sufficient evidence that the new program actually would…
(-Harm) … harm the environment,
(-Profit) … increase profits,
(-Help) … help the environment.
(Rel) The chairman thinks that the vice-president is generally reliable.
(Trust) The chairman thinks that the vice-president is generally trustworthy.
(Right) The chairman thought that the vice-president was probably right.

Participants specified their level of agreement using a seven-point Likert scale ranging from -3 to+3 (-3 was labeled as 'strongly disagree' and+3 as 'strongly agree').

## Appendix 2 Tables

**Table 7** Correlation between the attributions of knowledge and the attributions of belief (BHarm/BHelp), (consequence-) awareness (AwHarm/AwHelp), evidence (EvHarm/EvHelp), and justification (JuHarm/JuHelp)

| Outcome | Variable | *r* (KHarm) | *p* | *r* (KHelp) | *p* |
|---|---|---|---|---|---|
| Harm | Belief | 0.64 | < 0.001 | 0.58 | < 0.001 |
| | Awareness | 0.74 | < 0.001 | 0.63 | < 0.001 |
| | Evidence | 0.59 | < 0.001 | 0.49 | < 0.001 |
| | Justification | 0.65 | < 0.001 | 0.56 | < 0.001 |
| Help | Belief | 0.66 | < 0.001 | 0.71 | < 0.001 |
| | Awareness | 0.59 | < 0.001 | 0.66 | < 0.001 |
| | Evidence | 0.54 | < 0.001 | 0.61 | < 0.001 |
| | Justification | 0.54 | < 0.001 | 0.65 | < 0.001 |

**Table 8** Distribution of justifications selected by the participants depending on the (trichotomized) response to (KHarm)

| Outcome | KHarm | Yes | | No | | Neither | |
|---|---|---|---|---|---|---|---|
| | | % | *n* | % | *n* | % | *n* |
| Harm | Q | 50.00 | 66 | 18.64 | 11 | 23.53 | 4 |
| | A | 5.30 | 7 | 0 | 0 | 11.76 | 2 |
| | E | 12.12 | 16 | 28.81 | 17 | 29.41 | 5 |
| | B | 26.52 | 35 | 42.37 | 25 | 29.41 | 5 |
| | Ot | 6.06 | 8 | 10.17 | 6 | 5.88 | 1 |
| Help | Q | 40.96 | 34 | 11.82 | 13 | 21.43 | 6 |
| | A | 9.64 | 8 | 21.82 | 24 | 14.29 | 4 |
| | E | 13.25 | 11 | 12.73 | 14 | 39.29 | 11 |
| | B | 32.53 | 27 | 43.64 | 48 | 3.57 | 1 |
| | Ot | 3.61 | 3 | 10.00 | 11 | 21.43 | 6 |

**Acknowledgements** We would like to thank two anonymous reviewers for making many useful comments that helped to improve the paper.

**Funding** The project has been financed by a grant (2018/29/B/HS1/02861) from the National Science Centre, Poland.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

Alfano, M., J.R. Beebe, and B. Robinson. 2012. The Centrality of Belief and Reflection in Knobe-Effect Cases: A Unified Account of the Data. *The Monist* 95 (2): 264–289. https://doi.org/10.5840/monist201295215.

Beebe, J.R. 2013. A Knobe Effect for Belief Ascriptions. *Review of Philosophy and Psychology* 4: 235–258. https://doi.org/10.1007/s13164-013-0132-9.

Beebe, J.R. 2016. Do Bad People Know More? Interactions between Attributions of Knowledge and Blame. *Synthese* 193: 2633–2657. https://doi.org/10.1007/s11229-015-0872-4.

Beebe, J.R., and W. Buckwalter. 2010. The Epistemic Side-Effect Effect. *Mind and Language* 25: 474–498. https://doi.org/10.1111/j.1468-0017.2010.01398.x.

Beebe, J.R., and M. Jensen. 2012. Surprising Connections betweeen Knowledge and Action: The Robustness of the Epistemic Side-Effect Effect. *Philosophical Psychology* 25 (5): 689–715. https://doi.org/10.1080/09515089.2011.622439.

Beebe, J.R., and J. Shea. 2013. Gettierized Knobe Effects. *Episteme* 10 (3): 219–240. https://doi.org/10.1017/epi.2013.23.

Bratman, M.E. 1987. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press.

Buckwalter, W. 2014. Gettier Made ESEE. *Philosophical Psychology* 27 (3): 368–383. https://doi.org/10.1080/09515089.2012.730965.

Buckwalter, W., D. Rose, and J. Turri. 2015. Belief through Thick and Thin. *Nous* 49–4: 748–775. https://doi.org/10.1111/nous.12048.

Butler, R.J. 1978. Report on Analysis "Problem" No. 16. *Analysis* 38 (3): 113–114. https://doi.org/10.2307/3327843.

Dalbauer, N., and A. Hergovich. 2013. Is What Is Worse More Likely? The Probabilistic Explanation of the Epistemic Side-Effect Effect. *Review of Philosophy and Psychology* 4: 639–657. https://doi.org/10.1007/s13164-013-0156-1.

Grice, P. 1989. *Studies in the Way of Words*. Cambridge: Harvard University Press.

Harman, G. 1976. Practical Reasoning. *The Review of Metaphysics* 29 (3): 431–463.

Hindriks, F. 2008. Intentional Action and the Praise-Blame Asymmetry. *The Philosophical Quarterly* 58: 630–641. https://doi.org/10.1111/j.1467-9213.2007.551.x.

Hindriks, F. 2019. Explanatory Unification in Experimental Philosophy: Let's Keep It Real. *Review of Philosophy and Psychology* 10: 219–242. https://doi.org/10.1007/s13164-018-0397-0.

Knobe, J. 2003a. Intentional Action and Side Effects in Ordinary Language. *Analysis* 63 (3): 190–194. https://doi.org/10.1093/analys/63.3.190.

Knobe, J. 2003b. Intentional Action in Folk Psychology: An Experimental Investigation. *Philosophical Psychology* 33: 309–324. https://doi.org/10.1080/09515080307771.

Knobe, J. 2004. Intention, Intentional Action and Moral Considerations. *Analysis* 64: 181–187. https://doi.org/10.1093/analys/64.2.181.

Knobe, J. 2007. Reason Explanation in Folk Psychology. *Midwest Studies in Philosophy* 31: 90–106. https://doi.org/10.1111/j.1475-4975.2007.00146.x.

Knobe, J. 2010. Person as Scientist, Person as Moralist. *Behavioral and Brain Sciences* 33: 315–329. https://doi.org/10.1017/S0140525X10000907.

Knobe, J., and G. Mendlow. 2004. The Good, the Bad and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology. *Journal of Theoretical and Philosophical Psychology* 24: 252–258. https://doi.org/10.1037/h0091246.

Murray, D., J. Sytsma, and J. Livengood. 2013. God Knows (But does God Believe?). *Philosophical Studies* 166–1: 83–107. https://doi.org/10.1007/s11098-012-0022-5.

Myers-Schulz, B., and E. Schwitzgebel. 2013. Knowing that P without Believing that P. *Nous* 47–2: 371–348. https://doi.org/10.1111/nous.12022.

Paprzycka, K. 2015. The Omissions Account of the Knobe Effect and the Asymmetry Challenge. *Mind and Language* 30 (5): 550–571. https://doi.org/10.1111/mila.1209.

Paprzycka, K. 2016. Intention, Knowledge, and Disregard for Norms. In *Uncovering Facts and Values*, vol. 107, ed. A. Kuźniar and J. Odrowąż-Sypniewska, 204–233. Leiden: Brill | Rodopi. https://doi.org/10.1163/9789004312654_015.

Paprzycka-Hausman, K. 2020. Knowledge of Consequences: An Explanation of the Epistemic Side-Effect Effect. *Synthese* 197: 5457–5490. https://doi.org/10.1007/s11229-018-01973-1.

Pettit, D., and J. Knobe. 2009. The Pervasive Impact of Moral Judgment. *Mind and Language* 24: 586–604. https://doi.org/10.1111/j.1468-0017.2009.01375.x.

Robinson, B., P. Stey, and M. Alfano. 2015. Reversing the Side-Effect Effect: The Power of Salient Norms. *Philosophical Studies* 172: 177–206. https://doi.org/10.1007/s11098-014-0283-2.

Schaffer, J., and J. Knobe. 2012. Contrastive Knowledge Surveyed. *Nous* 46 (4): 675–708. https://doi.org/10.1111/j.1468-0068.2010.00795.x.

Steiger, J.H. 1980. Tests for Comparing Elements of a Correlation Matrix. *Psychological Bulletin* 87 (2): 245–251. https://doi.org/10.1037/0033-2909.87.2.245.

Turri, J. 2014. The Problem of ESEE Knowledge. *Ergo* 1 (4): 101–127. https://doi.org/10.2139/ssrn.3649796.

Yuan, Y., and M. Kim. 2021. Cross-Cultural Convergence of Knowledge Attribution in East Asia and the US. *Review of Philosophy and Psychology:* 1–28. https://doi.org/10.1007/s13164-021-00523-y.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.