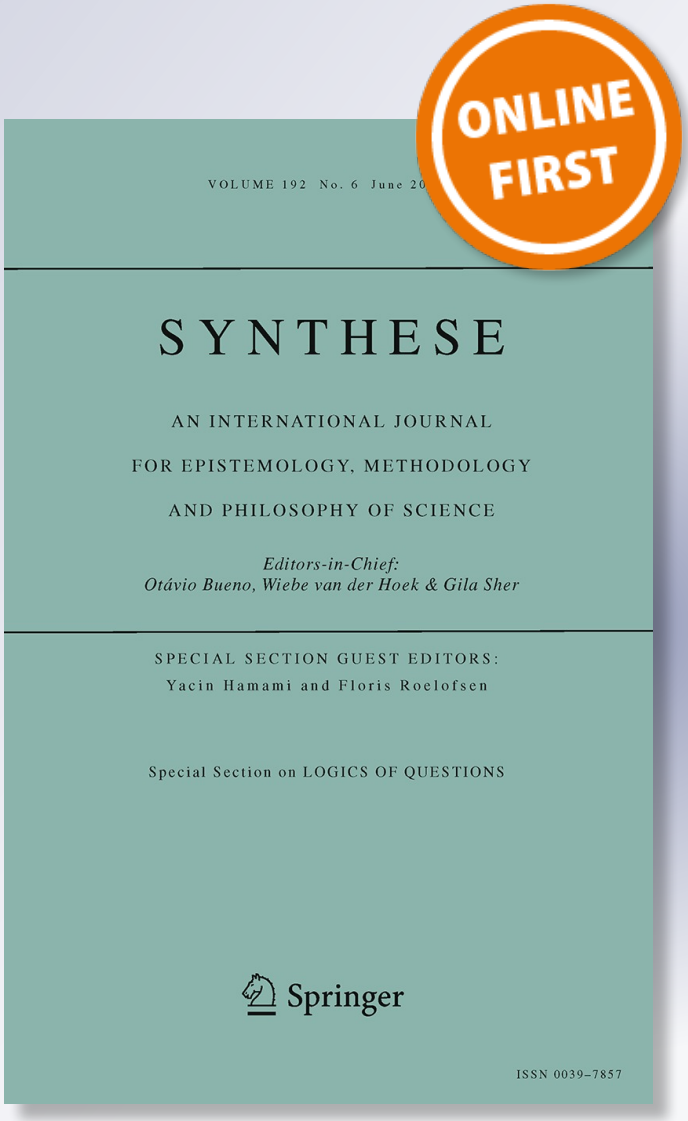
# *Do bad people know more? Interactions between attributions of knowledge and blame*

**James R. Beebe**

**Synthese**
An International Journal for Epistemology, Methodology and Philosophy of Science

ISSN 0039-7857

Synthese
DOI 10.1007/s11229-015-0872-4




Springer

Synthese
DOI 10.1007/s11229-015-0872-4

# Do bad people know more? Interactions between attributions of knowledge and blame

**James R. Beebe**[1]

Received: 27 July 2014 / Accepted: 22 August 2015


**Abstract** A central topic in experimental epistemology has been the ways that non-epistemic evaluations of an agent's actions can affect whether the agent is taken to have certain kinds of knowledge. Several scholars (e.g., Beebe and Buckwalter Mind Lang 25:474–98; 2010; Beebe and Jensen Philosophical Psychology 25:689–715, 2012; Schaffer and Knobe Noûs 46:675–708, 2012; Beebe and Shea Episteme 10:219–40, 2013; Buckwalter Philosophical Psychology 27:368–83, 2014; Turri Ergo 1:101–127, 2014) have found that the positive or negative valence of an action can influence attributions of knowledge to the agent. These evaluative effects on knowledge attributions are commonly seen as performance errors, failing to reflect individuals' genuine conceptual competence with KNOWS. In the present article, I report the results of a series of studies designed to test the leading version of this view, which appeals to the allegedly distorting influence of individuals' motivation to blame. I argue that the data pose significant challenges to such a view.

**Keywords** Experimental epistemology · Folk epistemology · Knowledge · Blame · Knobe effect · Epistemic side-effect effect

## 1 Introduction

Although researchers in experimental epistemology often examine the relative contributions of traditional factors like justification, truth, and belief to folk knowledge attributions (e.g., Starmans and Friedman 2012; Nagel et al. 2013; Myers-Schulz and Schwitzgebel 2013; Sackris and Beebe 2014), the most prominent research programs

✉ James R. Beebe
jbeebe2@buffalo.edu

[1] Department of Philosophy, University at Buffalo, Buffalo, NY 14260, USA

Published online: 04 September 2015

in experimental epistemology have focused on non-epistemic factors that influence epistemic judgments. For example, in the widely discussed findings of Weinberg et al. (2001), American college students of European ancestry were significantly less likely than American college students of East Asian (i.e., Korean, Japanese and Chinese) or South Asian (i.e., Indian, Pakistani, Bangladeshi) descent to think the protagonist in a Gettier case 'really knows' rather than 'only believes' a certain fact. Weinberg et al. (2001) also reported that high socioeconomic status participants were significantly more likely than low socioeconomic status participants to deny knowledge to agents in cases where skeptical questions had been raised. However, most philosophers seem unaware that Weinberg, Nichols, and Stich's cross-cultural result has failed to replicate and that they no longer take their socioeconomic finding to be reliable.[1] Pinillos and Simpson (2014) a second set of non-epistemic factors that experimental epistemologists have studied are the stakes or practical costs facing a believer. Some mainstream epistemologists (e.g., Stanley 2005; DeRose 1992, 1995, 2011; Fantl and McGrath 2002, 2007, 2009) have argued on philosophical grounds that the costs to a believer of a belief's being wrong should affect whether the belief counts as knowledge. Although initial attempts to find evidence for this view among the folk failed (Buckwalter 2010; Feltz and Zarpentine 2010; May et al. 2010; Phelan 2013), Ángel Pinillos (2012; Pinillos and Simpson 2014) and Sripada and Stanley (2012) have recently produced empirical evidence suggesting that folk knowledge attributions may well be affected by considerations of practical costs.

The present article focuses on a third set of non-epistemic factors that may play an important role in folk epistemology: moral and other broadly evaluative judgments of the goodness or badness of a believer's action. Building upon the work of Knobe (2003a, b, 2004), Beebe and Buckwalter (2010) found that individuals were not equally likely to think that the protagonist in the 'help' and the 'harm' versions of the following vignette knew that his actions would result in the predicted side-effects:

> *Environment*. 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, *and/but* it will also *help/harm* the environment." The chairman of the board answered, "I don't care at all about *helping/harming* the environment. I just want to make as much profit as I can. Let's start the new program." They started the new program. Sure enough, the environment was *helped/harmed*.

In response to the question, 'Did the chairman know that the new program would help (or harm) the environment?' Beebe and Buckwalter found that participants were significantly more likely to indicate that the chairman in the harm condition had knowledge ($M = 2.25$, on a scale from $-3$ to 3) than they were to think that the chairman in the help condition did ($M = 0.91$). This 'epistemic side-effect effect' (hereafter 'ESEE') seems surprising, given that the chairman's evidence appears to be the same in both cases.

[1] Cf. Cullen (2010), Nagel (2012), and Beebe (2012) for overviews of some of the most important criticisms of this work.

The ESEE could lead to profound changes in our understanding of the nature of knowledge, if it were determined that individuals are accurately displaying their conceptual competence with KNOWS when they allow the valence of an action undertaken in light of the belief that $p$ to affect their judgments about whether someone knows that $p$. Epistemologists would then have one more non-epistemic factor—in addition to stakes—to consider incorporating into their analyses of knowledge.

Many mainstream and experimental epistemologists, however, do not think that the ESEE should cause us to change our understanding of knowledge because they think it represents a performance error. It is also common to view the Knobe effect—i.e., Knobe (2003a, b, 2004) finding that individuals are more likely to say that a bad side-effect was brought about intentionally than a good one—as a performance error as well. Because something like the Knobe effect shows up in a variety of domains, it is reasonable to ask whether the same underlying factors might be responsible for the various versions of the effect. It is thus unsurprising that the most common explanation of the ESEE as a performance error draws upon prominent distortion-based accounts of the original Knobe effect. Adams and Steadman (2004a, b), Malle and Nelson (2003), Nadelhoffer (2004a, 2006), Alicke (2000, 2008), and Alicke and Rose (2010) all suggested that individuals' motivation to blame the central protagonists in Knobe effect cases had a distorting effect upon their intentionality attributions.[2] Following suit, some authors (e.g., Schaffer and Knobe 2012; Buckwalter 2014; Turri 2014) have considered the possibility that blame might be responsible for the ESEE as well. As we will see, however, blame-based explanations of the ESEE in the published literature are rather like skeptical hypotheses, in that discussions of them have tended to be authored by those who reject them rather than by those who endorse them.

Some blame-based explanations of the Knobe effect focus on pragmatic features of language use. For example, Adams and Steadman (2004a, b) note that an assertion of 'You did that intentionally' can serve to implicate blame. If denying that the chairman intentionally harmed the environment would generate the implicature that the chairman is not to blame for his action, this might increase intentionality attributions in the harm condition, even if the semantic content of 'intentionally' does not include blame as a component. One could follow Adams and Steadman and claim that since an assertion of 'The chairman knew the environment would be harmed' might also implicate that he is to blame for his action (while a denial of this claim might implicate that he was not to blame) we should expect knowledge attributions to be higher for blameworthy agents.

Most of the blame-based explanations of the Knobe effect are based upon a model of blame processing developed by Alicke (2000, pp. 556–558), who writes:

> Spontaneous evaluations encourage a blame-validation mode of processing in which evidence concerning the event is reviewed in a manner that favors ascribing blame to the person or persons who evoke the most negative affect or whose behavior confirms unfavorable expectations…. When a blame-validation

[2] Despite the fact that the blame hypothesis has been attributed to Nadelhoffer (2004a, 2006)—in large part because of the way his views have been discussed by Knobe (2007, 2010); Knobe and Mendlow (2004)—Nadelhoffer (2004b) maintains that more than blame is at work in Knobe effect cases.

> mode is engaged, observers review structural linkage evidence [i.e., evidence of connections between the mental states of an agent, the agent's behavior, and consequences of the action] in a biased manner by exaggerating the actor's volitional or causal control, by lowering their evidential standards for blame, or by seeking information to support their blame attribution.[3]

In other words, when a negative outcome is observed, a blameworthy agent is sought, a search for evidence to support ascriptions of blame is commenced, and the search process is carried out in a biased rather than an impartial manner. If Alicke is correct in thinking that a motivation to blame leads individuals to lower their evidential standards for blame, it would be surprising if they did not also lower their evidential standards for attributing knowledge. On this view, the tendency toward distorted folk psychological attributions does not occur when dealing with blameless individuals. Thus, the folk psychological attributions made to the chairman in the 'help' version of Environment should be viewed as free from the bias in question, although they may well be influenced by other biases.

In what follows, I report the results of seven new empirical studies that raise significant challenges to blame-based explanations of the ESEE. To fix ideas, the version of 'the blame hypothesis' that I will examine is comprised of the following claims:

(B1) The fact that individuals are more inclined to attribute knowledge to agents who perform negatively valenced actions than to agents who perform neutral or positively valenced actions represents a performance error.

(B2) The performance error consists in participants' increased tendency to attribute knowledge to agents who perform negatively valenced actions; participants' lesser inclination to attribute knowledge to agents who perform neutral or positively valenced actions does not represent an error of the kind in question.

(B3) The performance error is due to the undue influence of a motivation to blame agents who perform negatively valenced actions.

The blame hypothesis has the virtue of preserving deeply entrenched views about the 'epistemic purity' of knowledge. If evaluative effects on knowledge attributions represent errors, then our analyses of knowledge do not need to be changed to allow factors other than evidence or reliability to be what make the difference between knowledge and mere true belief.[4]

The few studies of blame-based explanations of the ESEE that have been published are largely unfavorable to the blame hypothesis. After reviewing existing studies in Sect. 2, I report the results of seven new studies in Sects. 3 through 9 that examine the blame hypothesis in further depth. I argue that these new data pose significant challenges to the blame hypothesis, rendering it unlikely to be true. In the con-

[3] Alicke (2000, p. 568) explains, "Blame-validation processing refers to observers' proclivity to favor blame versus nonblame explanations for harmful events and to de-emphasize mitigating circumstances…. Blame validation processing is encouraged by the tendency to view people rather than the environment as the prepotent controlling forces behind harmful events."

[4] Cf. Fantl and McGrath (2002, 2007, 2009) for discussion of epistemic purism and its rivals.

cluding section of the paper, I briefly suggest how the data on the ESEE might be better explained by an appeal to decision-theoretic features of agents' situations. The proposed explanation has the consequence that evaluative effects on knowledge attributions are not performance errors or biases after all.

## 2 Existing data

The environment cases were originally formulated by Knobe (2003a) to investigate whether the goodness or badness of a side-effect action might influence individuals' inclination to attribute intentionality to agents. Knobe found that 23 % of participants who read the help version thought the chairman intentionally helped the environment, whereas 82 harmed it. Knobe (2003a) also asked participants to rate on a scale from 0 to 6 how much praise they thought the helpful chairman deserved for what he did and how much blame they thought the harmful chairman deserved. Combining praise and blame ratings from the two environment cases and another pair of structurally similar cases, Knobe found that participants in the conditions where harm was brought about gave an average blame rating of 4.8, while participants in the help conditions gave an average praise rating of 1.4. Because these ratings correlate with participants' attributions of intentional action, Knobe (2003a, p. 193) suggests:

> And this asymmetry in people's assignment of praise and blame may be at the root of the corresponding asymmetry in people's application of the concept *intentional*: namely, that they seem considerably more willing to say that a side effect was brought about intentionally when they regard that side effect as bad than when they regard it as good.

As this quotation reveals, Knobe did not initially draw a clear distinction between the blameworthiness of the agent who brought about the bad side-effect and the badness of the side-effect itself. Once this distinction was highlighted by Nadelhoffer (2004a), Knobe (Knobe and Mendlow 2004; Knobe 2006) quickly committed himself to the idea that it was the badness of harmful side-effects rather than the blameworthiness of agents who bring them about that was responsible for generating the asymmetric pattern of intentionality attributions. However, Knobe never made it clear whether the contrast he was attempting to draw between badness and blame was supposed to map on to a contrast between consequentialist and deontological moral considerations or whether it should be understood in some other sense. Before addressing these questions, he moved on to consider other possible explanations of the data.[5] Nonetheless, the praise and blame ratings that Knobe obtained are consistent with a version of the blame hypothesis.

Knobe (2007) attempted to disconfirm blame-based accounts of the Knobe effect by asking participants to make judgments about the following vignette pair:

[5] For Knobe's most recent ruminations on the processes underlying the Knobe effect, cf. Pettit and Knobe (2009) and Knobe (2010). I will not examine in any detail the suggestions he makes about 'pro-attitudes with shifting default values' because knowledge is not obviously a pro-attitude, and it is not clear what the shifting defaults would be in the relevant epistemological cases.

> *Nazi*. In Nazi Germany, there was a law called the 'racial identification law.' The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps. Shortly after this law was passed, the CEO of a small corporation decided to make certain organizational changes. The vice-president of the corporation said: "By making those changes, you'll definitely be increasing our profits. But you'll also be *fulfilling/violating* the requirements of the racial identification law." The CEO said: "I don't care one bit about that. All I care about is making as much profit as I can. Let's make those organizational changes!" As soon as the CEO gave this order, the corporation began making the organizational changes.

Knobe asked participants whether the CEO intentionally fulfilled or violated the requirements of the law and found that 30 % of participants thought he intentionally fulfilled the law, while 81 % thought he intentionally violated the law. Knobe hypothesized that, while individuals may not consciously believe that the law-violating CEO's action was bad, nonconscious evaluative processes will pick up on the fact that he broke a law, classify this as bad, and modulate folk psychological attributions accordingly. That is, he conjectured that participants would treat the law-violating CEO like the environment-harming chairman. Beebe and Jensen (2012) used the same Nazi vignettes but asked participants whether they agreed that the CEO knew the organizational changes would fulfill or violate the requirements of the law. They found significantly more agreement in the 'violate' condition than in the 'fulfill' condition. These results accord with Knobe's predictions.

Knobe (2007) contends that data from the Nazi case pose a serious challenge for blame-based accounts of the Knobe effect. According to Knobe, the CEO in the violate condition did something that was bad in one sense but that was blameless overall, and thus the asymmetric attributions of intentionality he observed run contrary to what blame-based explanations would predict. In the language of dual-process approaches in psychology[6], Knobe suggests that while individuals' more conscious, deliberate, and reflective (System-2) cognitive processes recognize that the CEO is morally blameless for violating an unjust law, more automatic (System-1) processes are triggered by the badness of violating a law. Although violating an unjust law can be a good thing, the background expectations of the average American participant (who lives in a nation committed to the rule of law) are such that violating a law is generally bad. Participants' automatic responses to the badness of the Nazi CEO's action thus bias their attributions of intentionality and knowledge.

Pace Knobe, however, it does not seem that Knobe's data can be unproblematically interpreted as evidence against blame-based accounts. It is not clear that participants do in fact view the Nazi CEO's action as bad, and more importantly, if they did, it is not clear that they would understand the badness as being independent of considerations of blame. More importantly, however, advocates of blame-based accounts can and indeed have (Alicke 2000) appealed to the same kinds of unconscious (System-1) processes as Knobe. While upon reflection it is clear (to our System-2 processes)

[6] For accessible overviews of dual process theorizing in psychology, see Evans and Frankish (2009) and Kahneman (2011).

that the CEO who violates the racial identification law is blameless, this may not be clear to unconscious (System-1) processes associated with blame attribution. In other words, it may be automatic blame-attribution processes rather than automatic badness-apprehension processes that distort individuals' attributions of intentionality or knowledge. Knobe's brief study does nothing to show that his preferred explanation is the correct one.[7]

In another study of blame-based explanations of the Knobe effect, Liane Young and colleagues (Young et al. 2006) presented the Environment cases to subjects with deficits in emotional processing due to damage in the ventromedial prefrontal cortex (VMPC). Working on the assumption that the blame hypothesis must appeal to affect-driven blame attribution processes, they hypothesized that individuals with VMPC lesions should not display the Knobe effect. Young et al. found that subjects with VMPC lesions were nevertheless more inclined to judge that an action was intentional when it was morally bad than when it was morally good. Young et al. conclude that normal emotional processing does not seem to be responsible for Knobe effect and thus does not mediate the relationship between an action's moral status and its intentional status. Although this data is very interesting, it does not seem that proponents of the blame hypothesis need to be committed to the idea that blame attributions are affect-driven or, if they are, that they are subserved by the particular set of emotional processes that are compromised by VMPC lesions.

Another set of existing data that is relevant to our present concerns comes from Beebe and Jensen (2012), who modified Environment, so that the side-effects in question were movies that were made either better or worse from an artistic perspective (hereafter 'Movies') or sales in New Jersey that were either increased or decreased for the next quarter (hereafter 'Sales'). The basic structure of the cases remained the same. They found that the 'bad movies' and 'decreasing sales' CEOs were much more likely to be viewed as knowing the side-effects would occur than their 'better movies' and 'increasing sales' counterparts. Knobe and Mendlow (2004) asked participants whether the 'decreasing sales' CEO deserved any praise or blame for decreasing sales in New Jersey. Of the 75 % of subjects who thought the CEO intentionally decreased sales, only 5 % thought she deserved blame for this. 10 % thought she deserved praise, while 60 % said that she deserved neither praise nor blame. In a follow-up study, Phelan and Sarkissian (2008) asked participants to rate the degree of praise or blame the 'decreasing sales' CEO deserved for bringing about this decrease and to say whether they thought that decreasing sales was bad. In their study, 94.5 % of participants indicated that she deserved neither praise nor blame, and only 14 % of participants thought that decreasing sales was bad. These data do not accord well with the blame hypothesis.[8]

[7] There is also the issue that the Nazi case has a more complex structure than other vignettes employed in the Knobe effect and ESEE literatures. In Environment, for example, there is only one relevant norm that is violated, and it concerns how one's actions affect the environment. In Nazi, however, there are two salient norms that are violated: one that concerns obedience of local laws, and another that concerns being complicit in the deaths of innocent people. A proper understanding of the psychological factors at work in more complex cases like this cannot be attained by using only two experimental conditions.

[8] We should, however, keep in mind that proponents of the blame hypothesis can appeal to dual-process considerations and argue that participants' answers to explicit questions about blame might be generated

In another attempt to undermine blame-based explanations of the ESEE, Schaffer and Knobe (2012, p. 10), conducted a brief study in which they asked participants not about the knowledge of the chairman in environment but rather about the knowledge of a neutral, third party observer: "an environmentalist who knew that scientists were predicting helpful or harmful effects and then learned about the chairman's decision to go ahead with the program." When asked whether the environmentalist knew that the environment would be helped or harmed, significantly more participants indicated that the environmentalist knew that the environment would be harmed ($M = 4.8$, on a scale from 1 to 7) than indicated that he knew it would be helped ($M = 2.8$). In other words, they found the ESEE for a blameless observer. If this finding were to prove reliable, it would cast further doubt on the hypothesis that blame is what is driving increased knowledge attributions in negatively valenced Knobe effect and ESEE cases, since as the authors note, "there is no question of blaming the environmentalist for the outcome."

> Buckwalter (2014, p. 374, p. 376) obtained a similar finding using the following vignette:
> *Mayor*. The mayor of a small town is trying to decide whether or not to sign a new contract with a local corporation. The math is all very complex, but all his economic strategists think that there's a relatively good chance that one outcome is that it will *create/cut* jobs for workers in the community. The mayor says, "all I really care about is campaign contributions, not people's jobs, and I am sure to get millions from the corporation if I agree." [James the office secretary overheard everything, and is appalled by what the mayor said. Nonetheless, the mayor decides to sign the contract.] So, he decides to sign their contract. The corporation, however, didn't take any chances. They secretly switched the contract with a totally different one right before the mayor signed it. By changing all the fine print, in some cases the opposite of what the mayor thought he was signing, the corporation could be sure it got what it wanted. Sure enough, shortly after the mayor signed the contract, a number of members of the community *got/lost* jobs, and the mayor received a huge donation to his reelection campaign.

Mayor bears some similarity to Environment, except that its protagonist is Gettiered. In one pair of conditions, Buckwalter assigned participants to either the 'create jobs' or the 'cut jobs' condition and asked them whether the mayor knew that by signing the contract he would be creating or cutting jobs. In another pair of conditions, Buckwalter assigned participants to the create or cut jobs conditions but included the additional sentences in brackets above, which describe a third-party observer. For this second pair of conditions, participants were asked whether they agreed with the statement 'James the office secretary knew that members of the local community would *get/lose*

Footnote 8 continued
by controlled (System-2) processes that do not have access to the working of the automatic, unconscious (System-1) processes. In other words, the fact that participants said that a side-effect was not blameworthy does not mean that there was not some process in their minds that registered it as blameworthy. Despite these qualifications, the fact remains that one might have expected overactive blame attribution processes of the sort hypothesized by the blame account to have influenced participants' conscious (System-2) judgments about blame, but they do not appear to have done so.

**Table 1** Mean participant responses in Buckwalter (2014)

| | Mayor knew | Secretary knew |
|---|---|---|
| Create jobs | 4.11 | 3.95 |
| Cut jobs | 6.05 | 4.98 |

jobs.' Participant responses were recorded on a seven-point scale that ranged from 'Strongly disagree' to 'Strongly agree.' Buckwalter found that participants were still more inclined to attribute knowledge to the target subject, even when that subject was not responsible for the bad act in question (cf. Table 1).

Thus, Beebe and Jensen (2012), Schaffer and Knobe (2012), and Buckwalter (2014) all observed the ESEE with apparently blameless subjects. Neither this set of findings nor the praise and blame ratings obtained by Knobe (2003a), Knobe and Mendlow (2004), and Phelan and Sarkissian (2008) are favorable to the blame hypothesis.

On the other side of the coin, however, are the results of Beebe and Shea (2013). They found the ESEE in 'Knobified' Gettier cases only when the wrongdoing in question was committed by the character whose knowledge was the focus on the knowledge attribution question. For example, Beebe and Shea presented participants with different versions of the following, well-known case (due to Skyrms 1967) from the early days of the Gettier literature:

> *Pyromaniac*. A pyromaniac has just purchased a box of Sure-Fire Matches. He has done so many times before and has noted that they have always lit when struck unless they were wet. Furthermore, he knows that oxygen must be present for things to burn and that the observed regularity between the matches' being struck and their lighting is not a mere coincidence. After perceiving that the matches are dry and that there is plenty of oxygen present, he proceeds to strike one of the matches, confident that it will light. It does. Unbeknownst to the pyromaniac, however, the match happens to contain impurities that prevent it from lighting simply by being struck. What ignited the match was an extremely rare burst of cosmic radiation that happened to arrive at just the right place at the very moment the match was being struck.

A second version of this case was created by making the pyromaniac a disturbed criminal who wanted to use the Sure-Fire matches to burn down an orphanage in the middle of the night. An unGettiered control condition (i.e., a justified true belief case without Gettierization) was also created, where the match ignited for ordinary reasons without the intervention of cosmic radiation. Participant responses (registered on a five-point scale) were indistinguishable in the unGettiered control ($M = 3.95$) and Gettier plus wrongdoing ($M = 3.99$) conditions, but responses in both of these conditions were significantly higher than in the basic Gettier condition ($M = 3.23$).

However, when Beebe and Shea introduced elements of wrongdoing into traditional Gettier cases but made the person responsible for the wrongdoing someone other than the Gettiered subject, the ESEE disappeared. For example, they used the following, classic case from Harman (1973), p. 143:

> *Mail*. Gilbert's friend Donald tells Gilbert that he's going to Italy for the summer. In June Gilbert takes Donald to the airport and see him off. In July Donald decides

> to send Gilbert several letters informing him that he has gone to San Francisco. This is not true. Donald is simply trying to fool Gilbert. Donald sends the letters to another friend in San Francisco who is instructed to send them to Gilbert one at a time, as if they were sent from Donald, complete with a San Francisco postmark. Gilbert has not read any of these letters because he has been out of town. When he returns home, Gilbert's mail has piled up. Standing before a pile of unopened mail that includes two of Donald's phony letters, Gilbert still believes that Donald is in Italy. He is right. Donald is in Italy.

A Gettier plus wrongdoing condition was constructed by making Donald someone who had abducted a young child and was sending fake postcards to throw law enforcement officials off his track. An unGettiered control condition was created by making Donald travel to Italy and send postcards from there. In contrast to the Pyromaniac results, Beebe and Shea found that participant responses in the Gettier ($M = 3.32$) and Gettier plus wrongdoing ($M = 3.15$) conditions did not differ significantly from each other but did differ significantly from those in the unGettiered control condition ($M = 4.57$). Thus, in Beebe and Shea (2013) study, negatively valenced actions increased knowledge ratings only when the would-be knowers were the ones performing those actions. These data are consistent with the ESEE being driven by a motivation to blame.

Thus, most of the existing data concerning evaluative effects on knowledge attributions are unfavorable to the blame hypothesis. However, the data are not univocal, and some of the studies are not very thorough. For example, Schaffer and Knobe (2012) only reported their finding in a footnote, used a sample size of 28, and did not reproduce the research materials they used. Therefore, in an attempt to shed additional light on the interaction between attributions of blame and knowledge and to help settle the issue of the merits of the blame hypothesis, I performed the experiments described in the sections below.

## 3 Study 1

As we saw above, Knobe (2007) contends that data from the Nazi CEO case challenge blame-based accounts of the Knobe effect because in violating the racial identification law, the CEO performed an action that was blameless and indeed praiseworthy. More of Knobe's (2007) participants attributed intentionality to the law-violating CEO than to the law-fulfilling one, and more of Beebe and Jensen's (2012) participants attributed knowledge to the law-violating CEO. However, neither Knobe nor Beebe and Jensen obtained ratings of praise or blame from their participants. So, in order to test whether participants would naturally view the cases as Knobe predicted, in Study 1 I presented 117 undergraduates from a large, public university in the northeastern United States (40 % female, average age = 19, predominantly Caucasian) with either the 'fulfill' or the 'violate' version of Nazi. Half of the participants in each of the conditions were asked to rate (on a scale from −3 to 3) how much praise the CEO deserved for "performing this action" and the other half was asked how much blame he deserved. I left the CEO's action unspecified, so that participants' ratings might be more likely

**Table 2** Mean praise and blame ratings for the fulfill and violate conditions of Nazi in Study 1

| | Praise | Blame |
|---|---|---|
| Fulfill | −1.57 | 1.14 |
| Violate | 1.60 | −1.40 |

to reflect an all things considered judgment. Mean praise and blame ratings from this between-subjects study are summarized in Table 2.

Participants indicated that the CEO deserved praise rather than blame for violating the law and blame rather than praise for fulfilling it. These findings confirm Knobe's predictions about how individuals will understand the case. As we noted above, proponents of the blame hypothesis can appeal to unconscious blame attribution processes. But they owe us an explanation of why these overactive blame attribution processes can ratchet up attributions of knowledge and intentional action in the 'violate' conditions without affecting participants' explicit attributions of blame.

## 4 Study 2

In Sect. 2 I described how Schaffer and Knobe (2012) and Buckwalter (2014) both found the ESEE for blameless, third-party observers in variations of the Environment case. These data represent a challenge for the blame hypothesis because the hypothesis states that overactive motivations to blame agents who perform bad actions lead individuals to attribute more knowledge to those agents than to agents who perform good or neutral actions. But the agents to whom knowledge was more strongly attributed in these studies were blameless. If the observed effect arises without the hypothesized cause, there may be reason to doubt the hypothesis is correct. To investigate the reliability of these results, in Study 2 I created the following vignettes by taking the Environment, Movies, Sales, and Nazi cases mentioned above and adding a third-party observer to each one:

> *3rd Party Environment*. The CEO of a large company was thinking of starting a new program. An environmentalist closely monitoring the situation learned that scientists were reporting that the new program would *help/harm* the environment. The environmentalist then learned that the CEO had decided to start the new program. After the program was started, sure enough, the environment was *helped/harmed*.
> *3rd Party Movies*. The CEO of a movie studio was thinking of implementing a new policy. A journalist closely following the story learned that leading film critics were reporting that the new policy would make the studio's *better/movies* worse from an artistic standpoint. The journalist then learned that the CEO had decided to implement the new policy. After the policy was implemented, sure enough, the movies were made *better/worse* from an artistic standpoint.
> *3rd Party Sales*. Susan, the president of a major computer corporation, was thinking of implementing a new corporate restructuring plan. Susan's assistant learned that corporation's budget office was reporting that the restructuring plan would

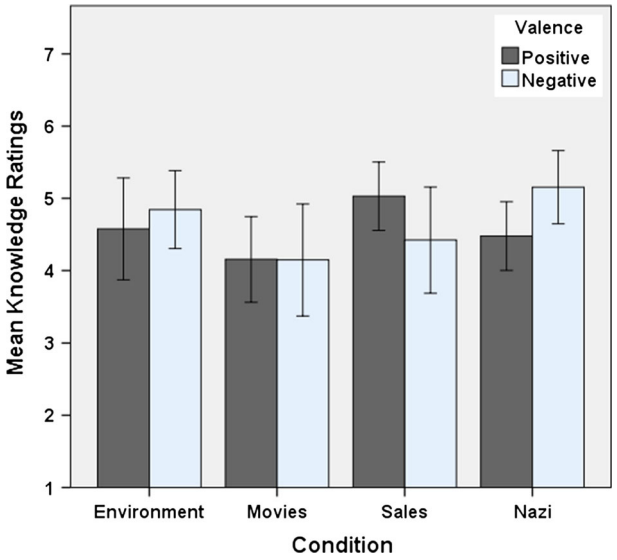



**Fig. 1** Mean knowledge ratings in the 3rd Party Environment (Help: $M = 4.58$; Harm: $M = 4.84$), 3rd Party Movies (Better: $M = 4.16$; Worse: $M = 4.15$), 3rd Party Sales (Increase: $M = 5.03$; Decrease: $M = 4.42$), and 3rd Party Nazi (Fulfill: $M = 4.48$; Violate: $M = 5.15$) conditions of Study 2. *Error bars* in all figures represent 95 % confidence intervals. (Although I calculate and report means on ordinal data throughout this paper, only nonparametric tests were used to determine statistical significance (with one exception in Study 7). Means do not figure in the calculations involved in nonparametric tests. I report means in tables and figures simply because there does not seem to be a more perspicuous summary statistic to use to present the data.)

*increase/decrease* sales in New Jersey in the following quarter. The assistant then learned that Susan had decided to implement the new plan. After the plan was implemented, sure enough, sales in New Jersey *increased/decreased* in the following quarter.

*3rd Party Nazi*. In Nazi Germany, there was a law called the "racial identification law." The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps. Shortly after this law was passed, the CEO of a small corporation was thinking about making certain organizational changes. The vice-president of the corporation learned that board members were reporting that the changes would *fulfill/violate* the requirements of the racial identification law. The vice-president then learned that the CEO had decided to make the changes.

In a between-subjects design, 273 participants (63% female, average age = 20, predominantly Caucasian) from a large, public university in the northeastern United States were asked whether they agreed that the neutral observer knew that the side-effect in question would occur. Participant ratings were recorded on a scale from 1 to 7 (ranging from Strongly Disagree to Strongly Agree) and are summarized in Fig. 1.

In none of the four pairs of matched cases was there a significant difference between the knowledge attributions that participants made to observers of good outcomes and

the attributions they made to observes of bad outcomes.[9] It is important to emphasize that Beebe and Buckwalter (2010) and Beebe and Jensen (2012) observed the ESEE with versions of each of these vignette pairs that did not involve neutral observers. All eight of the mean participant responses in Study 2 fell rather close to the neutral midpoint value of '4,' meaning the results were more similar to those obtained in the positive outcome conditions by Beebe and Buckwalter (2010) and Beebe and Jensen (2012) than to what they obtained in the negative outcome conditions. Thus, although Schaffer and Knobe (2012) observed the ESEE in their third-party version of Environment, I failed to do so. Furthermore, the results of Study 2 are consistent with what the blame hypothesis would predict.

## 5 Study 3

Because of the conflict between Buckwalter (2014) results and those of Study 2, I constructed a pair of follow-up experiments to adjudicate the issue. The most salient difference between the vignettes from Study 2 and Buckwalter's Mayor case (other than the rather complex structure of Mayor) is that Buckwalter's third-party observer evinces a strong emotional reaction to what he observes. James the office secretary is "appalled by what the mayor said." In order to see if controlling for this factor made any difference to participant knowledge attributions, in Study 3a I reran the 3rd Party Environment cases from Study 2 with online participants, but half the time the environmentalist was described as having a strong emotional reaction to what he observes. He was either "appalled that the CEO had decided to start the new program" when it was harmful or "overjoyed" when it was helpful. The unemotional versions of the vignettes were identical to the 3rd Party Environment cases above and thus represented an attempted replication. In each case, participants were asked whether they agreed that the neutral observer knew that the side-effect in question would occur.

160 participants (34 % female, average age = 34, predominantly Caucasian) were recruited from Amazon's Mechanical Turk (https://www.mturk.com/), each of whom had approval ratings of at least 98 % on more than 5000 Mturk tasks.[10] The results of Study 3a are summarized in Fig. 2.

The difference between the mean participant responses in each pair of cases was significantly different, with medium and small effect sizes, respectively.[11] Thus, the central finding of Study 2 failed to replicate, casting doubt on the challenge that Study 2 seemed to pose to (Schaffer and Knobe 2012) and (Buckwalter 2014) findings. The ESEE was found when a neutral observer was added to Environment and when that observer was not the person who committed the act of wrongdoing. Furthermore, giving the impartial observers strong emotional reactions only reduced but did not eliminate the ESEE.

[9] 3rd Party Environment: Mann-Whitney $U = 530.5$, $p > .05$. 3rd Party Movies: $U = 607.0$, $p > .05$. 3rd Party Sales: $U = 364.5$, $p > 05$. 3rd Party Nazi: $U = 426.5$, $p > .05$.

[10] Each was paid $0.35 for their work.

[11] 3rd Party Environment: $U = 471.5$, $p < .01$, $r = -.43$. 3rd Party Strong Reaction Environment: $U = 585.5$, $p < .05$, $r = -.26$.

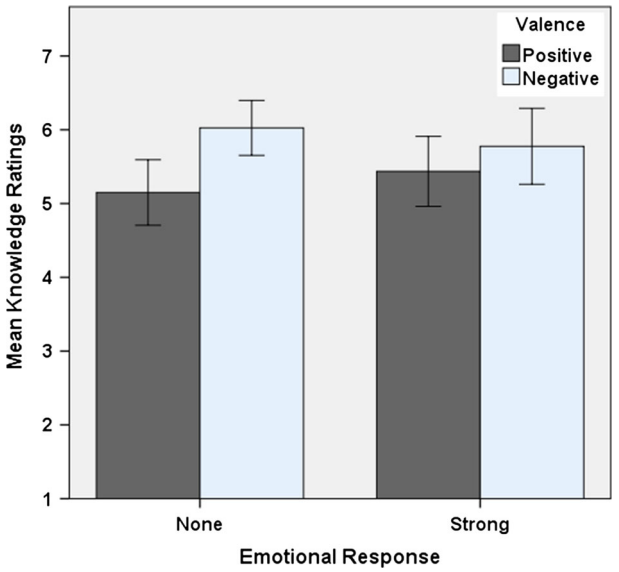



**Fig. 2** Mean knowledge ratings in the 3rd Party Environment (Help: $M = 5.1$; Harm: $M = 6.0$) and 3rd Party Strong Reaction Environment (Help: $M = 5.4$; Harm: $M = 5.8$) conditions of Study 3a

In Study 3b, I assigned some participants to read Buckwalter's original Mayor case (in which no observer was present) and his 3rd Party Mayor case (in which an observer has a strong emotional reaction). I then created a version of Buckwalter's third-party Mayor vignette in which the secretary simply "learns that the mayor decided to sign the contract." The 240 participants (50 % female, average age = 36, predominantly Caucasian) in this between-subjects study were workers from Amazon's Mechanical Turk with the same qualifications as above. The results of Study 3b are summarized in Fig. 3.

The mean participant responses in each pair of cases were significantly different, with medium to large effect sizes.[12] In other words, omitting the strong emotional reaction from Buckwalter's original 3rd Person Mayor case did not eliminate the ESEE. The results of Studies 3a and 3b, together with those of Schaffer and Knobe (2012) and Buckwalter (2014), do not bode well for the blame hypothesis. Even if the ESEE does not always manifest itself (as in Study 2), if taking the very cases that have demonstrated the ESEE and shifting the target of knowledge attributions to a blameless observer can still produce the ESEE, it does not seem plausible to claim that overactive blame attribution processes are responsible for driving the ESEE.

Above I noted that if the observed effect (increased knowledge attributions) can be observed in structurally similar cases without the hypothesized cause (overactive motivations to blame), there may be reason to doubt the causal story in question. However, a defender of the blame hypothesis might contend that overactive blame

[12] Mayor: $U = 428.5$, $p < .001$, $r = -.48$. 3rd Party Strong Reaction Mayor: $U = 408.5$, $p < .001$, $r = -.50$. 3rd Party No Reaction Mayor: $U = 504.0$, $p < .01$, $r = -.36$.

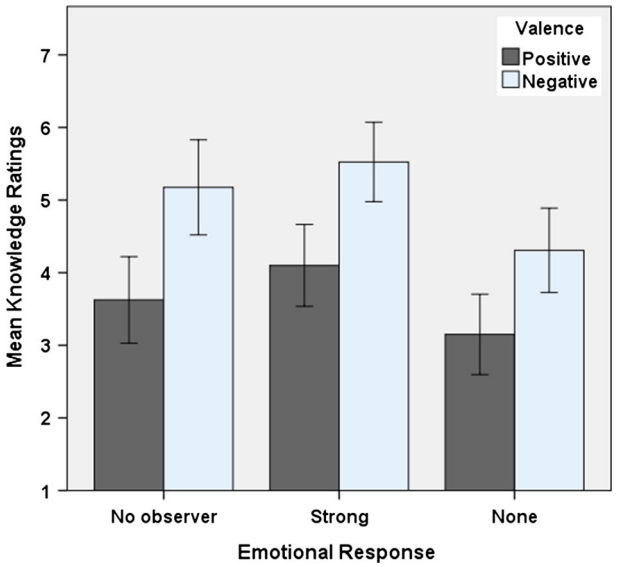



**Fig. 3** Mean participant responses to Mayor (Create jobs: $M = 3.6$; Cut jobs: $M = 5.2$), 3rd Party Strong Reaction Mayor (Create jobs: $M = 4.1$; Cut jobs: $M = 5.5$), and 3rd Party No Reaction Mayor (Create jobs: $M = 3.2$; Cut jobs: $M = 4.3$) cases in Study 3b

processes are nonetheless responsible for the data from Schaffer and Knobe (2012), Buckwalter (2014), and Studies 3a and 3b by arguing in the following fashion.[13] Individuals who attribute more knowledge to blameworthy protagonists in ESEE cases than blameless ones might be attributing knowledge to blameless observers as an indirect way of attributing knowledge to the blameworthy protagonists. Thus, blame could still be having an effect. Its influence would simply be indirect rather than direct. This possible explanation is not ruled out by the existing data on the ESEE. However, the more indirect the influence of blame attributions is hypothesized to be, and the more epicycles that must be added to make the account work, the less plausible it can seem.

## 6 Study 4

If a motivation to blame is responsible for the ESEE, one way to test this hypothesis is to transform the blameworthy chairman who harmed the environment into a more praiseworthy (or at least less blameworthy) character. Drawing upon an idea from Mele (2006), in Study 4a I changed the chairman's callous attitude, as expressed in "I don't care at all about harming the environment" and "I just want to make as much profit as I can" into the following, more admirable sentiment:

> *Reluctant Chairman*. I truly wish that I could make money for this company without harming the environment. Unfortunately, that seems to be impossible. Reluctantly, I'm instructing you to start the new program.

[13] Thanks to an anonymous reviewer for raising this point.

Although this chairman still harmed the environment, his expressed concern for the environment seems to make him a less despicable than the original chairman in the harm condition. In a between-subjects design, 116 undergraduates (52 % female, average age = 20, predominantly Caucasian) from a large, public university in the northeastern United States were asked whether the reluctant chairman knew that the new program would harm the environment. The mean response was 2.71 (on a scale from −3 to 3), which was significantly higher than the mean Beebe and Buckwalter obtained in the harm condition (2.25).[14] In other words, making the chairman less blameworthy in this case increased rather than decreased knowledge attributions to him.

In order to verify that the reluctant chairman was indeed viewed as less blameworthy, an additional 60 online participants (43 % female, average age = 38, predominantly Caucasian) from Amazon's Mechanical Turk were given both the original Environment case and the Reluctant Chairman case, with the former labeled 'Chairman1,' and the latter 'Chairman2.' Participants were then asked the following question:

> Please indicate which of the following statements you think best describes the blameworthiness of the two chairmans:
> Chairman1 is more blameworthy than Chairman2.
> Chairman2 is more blameworthy than Chairman1.
> Neither Chairman is more blameworthy than the other.

No participants thought that Chairman2 was more blameworthy, 23 thought that Chairman1 was more blameworthy, and 36 thought that neither was more blameworthy than the other.[15] Thus, the distribution of blame is shifted away from Chairman2.

Strictly speaking, the blame hypothesis does not predict that the only time we will observe increased knowledge attributions in ESEE cases is when blameworthy believers are involved. The blame hypothesis is thus compatible with there being other factors that can also lead to high knowledge attributions. However, the fact remains that the blame hypothesis does not seem able to explain the data of the present study. If, as I will argue below, there is a competing hypothesis that can explain both the original ESEE data and the results of Study 4, that hypothesis should be preferred over the blame hypothesis.

## 7 Study 5

Following another suggestion from Mele and Cushman (2007), I constructed a second pair of cases that featured an agent who displayed either callousness or concern about causing a side-effect that would be unpleasant to others:

> *Pond*. Al said to Ann: "If you fill in that pond in the empty lot next to your house, you're going to make the kids who look for frogs there sad." Ann replied: "*I don't care about making those kids sad / I like those kids, and I would definitely*

[14] $U = 18{,}383.5$, $p < .01$, $r = -.12$ (small effect size). The participants were undergraduates from the same university at which Beebe and Buckwalter (2010) and Beebe and Jensen (2012) conducted their studies.

[15] The qualifications and payment of these participants were the same as in the online studies above. An additional participant failed to answer the question correctly.

> *regret making them sad. The / But the* pond is a breeding ground for mosquitoes; and because I own the lot, I am responsible for it. It must be filled in." Ann filled in the pond, and, sure enough, the kids were sad.

In a between-subjects design, 235 undergraduates (54 % female, average age = 21, predominantly Caucasian) from a large, public university in the northeastern United Stated were asked whether they thought that Ann knew that filling in the lot would make the kids sad. Using a seven-point scale from −3 ('Strongly Disagree') to 3 ('Strongly Agree'), participants were somewhat more inclined to attribute knowledge to Ann when she regretted making the kids sad ($M = 1.65$) than when she said she didn't care about making them sad ($M = 1.41$), although this difference failed to be statistically significant.[16] Again, if blame were the key factor driving increased knowledge attributions in the harm condition of Environment, one might expect to see higher knowledge ratings when Ann says "I don't care about making those kids sad" than when she says "I like those kids, and I would definitely regret making them sad."

As in Study 4, an additional online 60 participants (40 % female, average age = 36, predominantly Caucasian) were also given both versions of Pond, with the one featuring Ann saying "I don't care" labeled 'Ann1,' and the one with her being more sympathetic labeled 'Ann2.' Participants were asked which Ann was more blameworthy. Only 2 thought that Ann2 was more blameworthy, with 24 attributing more blame to Ann1, and 34 saying that neither was more blameworthy than the other. Again, the blame hypothesis seems unable to explain high knowledge ratings when blameless agents are concerned.

## 8 Study 6

In the negatively valenced versions of the Environment, Movies, Sales, Nazi, and Pond cases above, the protagonist's primary action results in a side-effect that is bad in some sense. The environment is harmed, movies are made worse, sales are decreased, a law is violated, and children are made sad.[17] Importantly, the badness in each of these cases is inflicted by the protagonist upon someone or something else. Following a suggestion from Nadelhoffer (2004a, p. 209), I constructed the following pair of cases in which a blameless agent brings about a harmful side-effect upon himself:

> *Debate*. Steve and Jason are two friends who are competing in two different debate competitions. Jason decides to help Steve prepare for his debate. Ellen, a mutual friend, says, "Don't you realize that by spending so much time helping Steve with his debate you are *increasing/decreasing* your chances of winning your own debate?" Jason responds, "I don't care at all about that. I just want to help my friend!" Steve went on to win his debate competition because of Jason's help, but Jason did not win his.

[16] $U = 6508.0$, $p > .05$.

[17] I do not want to suggest that these outcomes are bad in an 'all things considered' sense—merely that they are bad in some respect.

When Jason says that he does not care about increasing or decreasing his own chances of winning, his lack of concern is admirable (especially when it decreases his chances), in contrast to the chairman's lack of concern about the environment. In a between-subjects design, 144 undergraduates (54 % female, average age = 21, predominantly Caucasian) from a large, public university in the northeastern United States were asked in a forced-choice format whether they thought that Jason knew or did not know he would be increasing or decreasing his chances of winning his own debate. 45 % of participants thought that Jason knew he would be increasing his chances, but 77 % of participants thought that he knew he would be decreasing his chances. This difference is statistically significant.[18]

The Reluctant Chairman, Pond, and Debate cases present indirect challenges to the blame hypothesis, insofar as it seems unable to explain increased knowledge ratings to blameless or conscientious agents. Nadelhoffer (2004a) has suggested that the basic blame hypothesis could be modified, so that either highly blameworthy or highly praiseworthy agents will be the target of distorted folk psychological attributions. An important difficulty for this proposal is that the blame hypothesis was already positing a great deal of distortion in the ordinary practice of folk psychological attribution. Now it will be positing even more distortion, this time along another dimension.

## 9 Study 7

In their seminal article on outcome bias, Baron and Hershey (1988) discovered that individuals rated the thinking behind certain decisions as better, rated the decision makers as more competent, and expressed greater willingness to let the decision makers make choices on their behalf when the outcomes were favorable than when they were unfavorable—even when the information available to decision makers remained constant.[19] Alicke and Davis (1989); Alicke et al. (1994), the leading thinker behind blame-based explanations of the Knobe effect, found that outcome bias tends to significantly increase ascriptions of blame to agents whose actions produce negative consequences. Alicke (2008) also notes that participants in the harm condition of Environment have a reason to blame the chairman before they are told how things turned out, since he expresses an attitude of reckless disregard toward potential environmental harm. However, Alicke hypothesizes that when participants read that the environment was in fact harmed by the chairman's decision to implement the new

[18] $\chi^2$ (1, $N = 144$) = 14.793, $p < .001$, Cramér's V = .32 (medium effect size).

[19] Outcome bias should be distinguished from hindsight bias, which occurs when reporting an outcome's occurrence unjustifiably increases its perceived probability of occurrence beyond what the prior information available would warrant. In his seminal research on hindsight bias, Fischhoff (1975); Fischhoff and Beyth (1975) found that individuals overestimated what they would have known without the outcome knowledge, as well as what others actually did know without this knowledge. Fischhoff (1975, 293) opines, "it appears that what passes for the wisdom of hindsight often contains heady doses of sophistry—that the perceived inevitability of reported outcomes is imposed upon, rather than legitimately inferred from, the available evidence." Although outcome information plays an important role in both biases, the distinction between them is that in hindsight bias outcome information skews the perceived (prior) probability of an event's occurring, while in outcome bias the outcome information distorts qualitative assessments of actions or decisions, independently of the probability of their occurrence.

**Table 3** Mean knowledge ratings in 'No Outcome Info.' conditions of Study 7, and the 'Outcome Info.' conditions of Beebe and Buckwalter (2010) and Beebe and Jensen (2012)

| | Good Outcome | | Bad Outcome | |
|---|---|---|---|---|
| | Outcome Info. | No Outcome Info. | Outcome Info. | No Outcome Info. |
| Environment | 0.91 | 0.77 | 2.25 | 1.16 |
| Movies | 0.92 | 0.50 | 1.50 | 0.45 |
| Sales | 0.80 | −0.43 | 1.50 | 0.53 |

program, outcome bias leads them to ascribe even more blame to the chairman, and this superabundance of blame motivation in turn causes them to attribute all manner of foresight and nefarious intentions to him.

Thus, according to Alicke, the blame hypothesis predicts that we should see a weaker tendency on the part of participants to attribute foresight and intention to the chairman in the harm condition, if they are not given any outcome information. Since the chairman's action in the help condition of Environment is neither terribly praiseworthy nor terribly blameworthy, the theory also predicts that the mechanisms involved in outcome bias will be less significantly engaged when outcome information is provided about this case and, therefore, that there should be less of an effect when such information is withheld.

In order to test these predictions, in Study 7 I presented participants with Beebe and Buckwalter (2010) and Beebe and Jensen (2012) environment, movies, and sales vignettes but eliminated the following final sentences from each of them, which contained information about how things turned out:

> *Environment*. Sure enough, the environment was helped/harmed.
> *Movies*. Sure enough, the policy made the movies better/worse from an artistic standpoint.
> *Sales*. Sure enough, sales in New Jersey increased/decreased in the following quarter.

In a between-subjects design, 329 undergraduate students (57 % female, average age = 20, predominantly Caucasian) from a large, public university in the northeastern United States were asked whether they agreed that the central protagonists in the vignettes knew that the side-effects in question would occur. Participants gave their responses on a scale ranging from −3 ('Strongly Disagree') to 3 ('Strongly Agree').

Results from Study 7 are summarized in the 'No Outcome Info.' cells of Table 3, along with data from Beebe and Buckwalter (2010) and Beebe and Jensen (2012) presented in the 'Outcome Info.' cells for the sake of comparison. The mean knowledge rating in each 'No Outcome Info.' cell was numerically lower than the mean in its 'Outcome Info.' counterpart, with the largest such differences being observed in pairs of bad outcome cases.

After transforming the combined data represented in Table 3 in order to reduce negative skewness in the original distributions[20], a three-way ANOVA (condition ×

[20] A constant was added to all values in the data set, in order to eliminate negative and zero values. Then the data were reflected, and the base 10 logarithm of each value was computed.

valence × outcome information) was performed, which revealed a main effect for all three variables.[21] In other words, there was a significant difference between the mean knowledge ratings in the good outcome conditions and the bad outcome conditions and between the outcome information and no outcome information conditions. There was even a significant difference between the knowledge ratings in the Environment, Movies, and Sales conditions. Post-hoc (Tukey's HSD) tests reveal that the mean ratings in the Environment conditions were significantly higher than those in Movies or Sales ($p < .001$), but that the ratings in the latter two sets of conditions did not differ significantly. In addition, there was a significant interaction effect between the valence and outcome information variables, meaning that eliminating outcome information had more of an effect on knowledge ratings in bad cases than in good cases.[22] There were no significant interactions between the condition and outcome information variables or between the condition and valence variables.[23]

Focusing on the results of Study 7 (i.e., only on the no outcome conditions), a two-way ANOVA (condition x valence) reveals a significant main effect for both condition and valence.[24] In other words, the mean knowledge ratings were significantly higher in the bad outcome conditions than in the good outcome conditions. Post-hoc (Tukey's HSD) tests again revealed that the mean knowledge ratings in the Environment conditions were significantly higher than those in Movies or Sales ($p < .001$), but that the ratings in the latter two sets of conditions did not differ significantly. Thus, although eliminating outcome information reduced the size of the ESEE, it did not make it disappear completely.

The most favorable outcome of Study 7 for the blame hypothesis would have been if the ESEE had disappeared when outcome information was eliminated. But this was not observed. Granted, proponents of the blame hypothesis can argue that protagonists who display indifference toward causing harm will still be viewed as blameworthy, even if participants are not told how things turned out, and that this perceived blame could still lead to increased knowledge attributions. But the fact remains that in a variety of cases we do not observe what the blame hypothesis would lead us to expect if it were true.

Even the fact that the eliminating outcome information decreased the size of the ESEE to some extent cannot be taken as confirmation of Alicke's (Alicke 2000, 2008; Alicke and Rose 2010) hypothesis about the operation of outcome bias because of the factivity of knowledge. Philosophers have traditionally viewed knowledge as being factive, meaning that if you know that $p$, $p$ must be true. Thus, the decreased size of the ESEE observed in Study 7 when outcome information was eliminated may simply stem from participants' tacit recognition that knowledge cannot be as confidently ascribed when the truth value of a belief is unknown, as compared to when it is known. Even

[21] Condition: $F(2, 1439) = 14.07$, $p < .001$, partial $\eta^2 = .02$. Valence: $F(1, 1439) = 36.61$, $p < .001$, partial $\eta^2 = .03$. Outcome information: $F(1, 1439) = 46.219$, $p < .001$, partial $\eta^2 = .03$.

[22] $F(1, 1439) = 3.93$, $p < .05$, partial $\eta^2 = .003$.

[23] Condition x outcome information: $F(2, 1439) = 1.02$, $p > .05$. Condition x valence: $F(2, 1439) = 2.48$, $p > .05$.

[24] Condition: $F(2, 323) = 7.58$, $p < .01$, partial $\eta^2 = .05$. Valence: $F(1, 323) = 6.16$, $p < .05$, partial $\eta^2 = .02$.

those who deny factivity should not deny that information about the truth value of a belief should be expected to have a significant effect on individuals' willingness to attribute knowledge.[25] Thus, the blame hypothesis once again fails to receive support from the available empirical information.

## 10 General discussion

Blame-based explanations of the Knobe effect have played a prominent role in debates about the folk concept of intentional action (e.g., Adams and Steadman 2004a, b; Malle and Nelson 2003; Nadelhoffer 2004a, 2006; Alicke 2000, 2008; Alicke and Rose 2010). Some experimental philosophers (e.g., Schaffer and Knobe 2012; Buckwalter 2014; Turri 2014) have considered the possibility that blame might be responsible for the ESEE as well. Despite the fact that some data are consistent with the blame hypothesis (e.g., Study 2 and Beebe and Shea 2013, on the whole it appears that its prospects as a successful explanation of the ESEE are rather dim. Praise and blame ratings of agents in ESEE cases (Study 1; Knobe and Mendlow 2004; Phelan and Sarkissian 2008) and knowledge attributions to neutral observers (Study 3, Schaffer and Knobe 2012; Buckwalter 2014) and blameless agents (Studies 4 through 6) suggest that if overactive blame attribution processes are having some influence, they are not the primary factor responsible for the large-scale patterns observed in ESEE cases.

In Sect. 1, I noted that the ESEE could lead to profound changes in our understanding of the nature of knowledge, if epistemologists became convinced that individuals are accurately reflecting their conceptual competence with KNOWS when they allow the valence of an action undertaken in light of the belief that $p$ to affect their judgments about whether someone knows that $p$. I speculate that most mainstream epistemologists would resist the idea that the ESEE does not represent a performance error. It should be noted that a successful refutation of the blame hypothesis should not be taken to imply that the ESEE is not the result of bias or distortion. There could be some factor (or set of factors) other than blame that responsible for skewing knowledge attributions in ESEE cases. However, the repudiation of the blame hypothesis does put pressure on proponents of distortion-based accounts to specify what those factors are. At present, there are no other viable candidates that have been put forward.

It is commonly thought that the only alternative to positing bias as the root cause of the ESEE is to accept that it results from folk conceptual competence with KNOWS.[26] However, I would like to briefly recommend a different kind of approach that does not fall neatly into either category, viz., the 'belief heuristic' approach articulated by Alfano et al. (2012). Beginning from the observation that some beliefs are more worth having than others, these authors argue that agents whose actions violate normative expectations are more likely to engage in deeper levels of reflection about the con-

[25] This suggestion gains support from recent experimental studies (e.g., Turri & Buckwalter forthcoming) that information about truth value is a significant factor in prompting ordinary individuals to attribute knowledge.

[26] Cf., e.g., Buckwalter (2014) and Turri (2014).

sequences of their actions than agents who do not violate such expectations.[27] The write:

> [T]rue beliefs to the effect that one is violating a norm are typically more valuable than true beliefs to the effect that one is conforming to a norm. One may be sanctioned for violating a norm, so forming a true belief about whether one has violated a norm (hence potentiating such a sanction) is valuable, regardless of whether one endorses the norm. The chairman in the HELP condition, for example, does not need to say to himself, "Wait! I need to stop and think carefully about whether helping the environment is something that I should be doing." In the HARM condition, however, an inner monologue like this might well be appropriate. (Alfano et al. 2012, p. 269)

Increased reflection about consequences makes it more likely that agents who violate normative expectations will form stronger beliefs and come to have knowledge about the consequences of their actions, and attributors will reasonably interpret them as having done so.[28]

Although Alfano, Beebe, and Robinson focus primarily on norm-violating behavior, it seems that a variety of features of an agent's decision situation that calls for greater reflection might lead to increased attributions of knowledge and belief to that agent. For example, just as the environment-harming chairman needed to reflect upon the consequences of his actions to a greater degree than his environment-helping counterpart, the reluctant chairman, both versions of Ann, and 'decreasing-chances' Jason all face a personal, practical cost that calls for attention and at least some degree of deliberation. On the belief heuristic account, participants rightly interpreted these agents as having formed stronger beliefs and as being more likely to have obtained knowledge about the occurrence of these costly side-effects because of this practical need for greater cognitive engagement. It seems that Alfano, Beebe, and Robinson's account can even explain data from the neutral, third-party observer cases of Schaffer and Knobe (2012), Buckwalter (2014), and Study 3. In general, people have more practical reasons to pay attention to the norm-violating behavior of others than to their norm-conforming behavior. This could again lead observers (a) to be more likely to form beliefs about the behavior, (b) to form stronger beliefs about the behavior, and (c) to be taken by others to have formed the beliefs described in (a) and (b).

This kind of explanation has the consequence that evaluative effects on knowledge attributions should not be viewed performance errors or biases, and it allows for this possibility without sacrificing the epistemic purity of knowledge. It is only when one assumes that the chairman in the help condition of Environment and the chairman in the harm condition are plausibly taken to have equally strong evidence for a proposition that they believe to same degree that one is led to think that the absence of performance

[27] The belief heuristic hypothesis shares some things in common with Uttich and Lombrozo's (2010) suggestion that norm-conforming behavior is generally less informative about the mental states underlying the behavior than norm-violating behavior. Because norms automatically provide reasons for acting in accord with them, Uttich and Lombrozo contend that norm-violating behavior requires and often points toward reasons for the norm violation. According to the belief heuristic hypothesis, an action's practical costs and benefits provide information about what those reasons are likely to be.

[28] Beebe's (2013) studies provide direct confirmation for this claim.

errors means the analysis of the folk conception of knowledge must be modified to incorporate new factors.[29] But if important features of the agents' psychological states in ESEE cases are not in fact viewed as being the same, recognizing this can make room to explore new kinds of explanations of the data. In sum, while further research needs to be done on the various ways that evaluative judgments of an agent's action can affect attributions of knowledge to that agent, it seems that factors other than blame need to be given serious consideration.

## References


Adams, F., & Steadman, A. (2004a). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, *64*, 173–181.

Adams, F., & Steadman, A. (2004b). Intentional action and moral considerations: Still pragmatic. *Analysis*, *64*, 268–276.

Alfano, M., Beebe, J. R., & Robinson, B. (2012). The centrality of belief and reflection in Knobe effect cases: A unified account of the data. *The Monist*, *95*, 264–89.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556–74.

Alicke, M. D. (2008). Blaming badly. *Journal of Cognition and Culture*, *8*, 179–86.

Alicke, M. D., & Davis, T. L. (1989). The role of a posteriori victim information in judgments of blame and sanction. *Journal of Experimental Social Psychology*, *25*, 362–77.

Alicke, M. D., Davis, T. L., & Pezzo, M. V. (1994). A posteriori adjustment of a priori decision criteria. *Social Cognition*, *12*, 281–308.

Alicke, M., & Rose, D. (2010). Culpable control or moral concepts? *Behavioral and Brain Sciences*, *33*, 330–1.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569–79.

Beebe, J. R. (2012). Experimental epistemology. In A. Cullison (Ed.), *Companion to Epistemology* (pp. 248–269). London: Continuum.

Beebe, J. R. (2013). A Knobe effect for belief ascriptions. *Review of Philosophy and Psychology*, *4*, 235–58.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language*, *25*, 474–98.

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

Beebe, J. R., & Shea, J. (2013). Gettierized Knobe effects. *Episteme*, *10*, 219–40.

Buckwalter, W. (2010). Knowledge isn't closed on Saturday: A study in ordinary language. *Review of Philosophy and Psychology*, *1*, 395–406.

Buckwalter, W. (2014). Gettier made ESEE. *Philosophical Psychology*, *27*, 368–83.

Cullen, S. (2010). Survey-driven romanticism. *Review of Philosophy and Psychology*, *1*(2), 275–296.

Dalbauer, N., & Hergovich, A. (2013). Is what is worse more likely? The probabilistic explanation of the epistemic side-effect effect. *Review of Philosophy and Psychology*, *4*, 639–657.

DeRose, K. (1992). Contextualism and knowledge attributions. *Philosophy and Phenomenological Research*, *52*(4), 913–929.

DeRose, K. (1995). Solving the skeptical problem. *Philosophical Review*, *104*(1), 1–52.

DeRose, K. (2011). *The case for contextualism: Knowledge, skepticism, and context* (Vol. 1). Oxford: Oxford University Press.

Evans, J., & Frankish, K. (2009). *In two minds: Dual processes and beyond*. Oxford: Oxford University Press.

Fantl, J., & McGrath, M. (2002). Evidence, pragmatics, and justification. *Philosophical Review*, *111*, 67–94.

Fantl, J., & McGrath, M. (2007). On pragmatic encroachment in epistemology. *Philosophy and Phenomenological Research*, *75*, 558–589.

Fantl, J., & McGrath, M. (2009). *Knowledge in an uncertain world*. Oxford: Oxford University Press.


[29] Cf. Dalbauer and Hergovich (2013) for an investigation of whether individuals view positively and negatively valenced outcomes as being equally likely.

Feltz, A., & Zarpentine, C. (2010). Do You know more when it matters less? *Philosophical Psychology*, *23*, 683–706.
Fischhoff, B. (1975). Hindsight ≠ foresight: the effect of outcome knowledge on judgement under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 288–99.
Fischhoff, B., & Beyth, R. (1975). I knew it would happen: Remembered probabilities of once-future things. *Organizational Behaviour and Human Performance*, *13*, 1–16.
Harman, G. (1973). *Thought*. Princeton: Princeton University Press.
Kahneman, D. (2011). *Thinking, fast and slow*. London: macmillan.
Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190–193.
Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*, 309–324.
Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, *64*, 181–187.
Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, *130*, 203–231.
Knobe, J. (2007). Reason explanation in folk psychology. *Midwest Studies in Philosophy*, *31*, 90–106.
Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*, 315–65.
Knobe, J., & Mendlow, G. (2004). The good, the bad, and the blameworthy: Understanding the role of evaluative considerations in folk psychology. *Journal of Theoretical and Philosophical Psychology*, *24*, 252–258.
Malle, B. F., & Nelson, S. E. (2003). Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law*, *21*, 563–80.
May, J., Sinnott-Armstrong, W., Hull, J. G., & Zimmerman, A. (2010). Practical interests, relevant alternatives, and knowledge attributions: An empirical study. *Review of Philosophy and Psychology*, *1*, 265–273.
Mele, A. (2006). The folk concept of intentional action: A commentary. *Journal of Cognition and Culture*, *6*, 277–90.
Mele, A., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, *31*, 184–201.
Myers-Schulz, B., & Schwitzgebel, E. (2013). Knowing that P without believing that P. *Noûs*, *47*, 371–84.
Nadelhoffer, T. (2004a). On praise, side effects, and folk ascriptions of intentionality. *The Journal of Theoretical and Philosophical Psychology*, *24*, 196–213.
Nadelhoffer, T. (2004b). Blame, badness, and intentional action: A reply to Knobe and Mendlow. *The Journal of Theoretical and Philosophical Psychology*, *24*, 259–269.
Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional action: Some problems for juror impartiality. *Philosophical Explorations*, *9*, 203–19.
Nagel, J. (2012). Intuitions about Gettier cases: A cross-cultural approach. Presented at the Eastern Division Meeting of the American Philosophical Association. Atlanta, GA, December 30, 2012.
Nagel, J., San Juan, V., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition*, *129*, 652–61.
Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, *24*, 586–604.
Phelan, M. (2013). Evidence that stakes don't matter to evidence. *Philosophical Psychology*, *4*, 1–25.
Phelan, M., & Sarkissian, H. (2008). The folk strike back: Or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, *138*, 291–98.
Pinillos, N. Á. (2012). Knowledge, experiments and practical interests. In J. Brown & M. Gerken (Eds.), *Knowledge Ascriptions* (pp. 192–219). Oxford: Oxford University Press.
Pinillos, N. Á., & Simpson, S. (2014). Experiment evidence supporting anti-intellectualism about knowledge. In J. R. Beebe (Ed.), *Advances in experimental epistemology* (pp. 9–43). London: Bloomsbury.
Sackris, D., & Beebe, J. R. (2014). Is justification necessary for knowledge? In J. R. Beebe (Ed.), *Advances in experimental epistemology* (pp. 175–192). London: Bloomsbury.
Schaffer, J., & Knobe, J. (2012). Contrastive knowledge surveyed. *Noûs*, *46*, 675–708.
Skyrms, B. (1967). The explication of 'X knows that P'. *Journal of Philosophy*, *64*, 373–89.
Sripada, C. S., & Stanley, J. (2012). Empirical tests of interest-relative invariantism. *Episteme*, *9*, 3–26.
Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford: Oxford University Press.
Starmans, C., & Friedman, O. (2012). The folk conception of knowledge. *Cognition*, *124*, 272–83.
Turri, J. (2014). The problem of ESEE knowledge. *Ergo*, *1*, 101–127.
Turri, J., & Buckwalter, W. (forthcoming). Descartes's schism, Locke's reunion: Completing the pragmatic turn in epistemology.

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

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, *29*, 429–60.

Young, L., Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. (2006). Does emotion mediate the effect of an action's moral status on its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture*, *6*, 265–278.