The Epistemic Side-Effect Effect

James Beebe and Wesley Buckwalter* (University at Buffalo)

Knobe (2003a, 2003b, 2004b) and others have demonstrated the surprising fact that the valence of a side-effect action can affect intuitions about whether that action was performed intentionally. Here we report the results of an experiment that extends these findings by testing for an analogous effect regarding knowledge attributions. Our results suggest that subjects are less likely to find that an agent knows an action will bring about a side-effect when the effect is good than when it is bad. It is further argued that these findings, while preliminary, have important implications for recent debates within epistemology about the relationship between knowledge and action.

Introduction

The 'side-effect effect' is one of the most widely discussed results of recent experimental philosophy. Using experimental survey methods, Joshua Knobe (2003a, 2003b, 2004b) and others have demonstrated that subjects are more inclined to say that an agent has intentionally performed a side-effect action if that action is bad than if it is good. This asymmetry in the intuitions of ordinary subjects challenges the bit of conventional wisdom in action theory that holds that assessments of intentionality are prior to assessments of blame.

We extend these findings from the intentionality literature by demonstrating an analogous effect involving the concept of knowledge. We show that subjects are more likely to say that agents know their actions will bring about certain side-effects, if the effects are bad than if they are good. Our results challenge the bit of conventional wisdom about knowledge attribution that holds that the moral significance of an action undertaken in light of or on the basis of a belief should have no effect on whether that belief counts as knowledge. Our results are thus relevant to recent debates within epistemology about whether non-epistemic factors affect knowledge attributions and about the relationship between knowledge and action.

In section 1 we introduce the original side-effect effect and the findings that constitute the epistemic side-effect effect. In the following section we briefly compare our findings to other results in the literature. In section 3 we survey the primary explanations of the side-effect effect that have been offered in the literature and suggest how extensions of those explanations might be applied to our own results. A final section (sec. 4) spells out some implications of our findings for epistemology and action theory and suggests avenues for future research.

1. Side-Effect Effects

Knobe (2003a) originally demonstrated the side-effect effect by asking subjects whether or not the CEO of a company brought about a certain side-effect intentionally. In one condition the side-effect was beneficial to the environment. In another condition, the side-effect was harmful. In both conditions, however, the chairman's evidence that the side-effect would certainly occur was identical. Surprisingly, Knobe found that 82% of subjects in the harm condition agreed that the chairman intentionally harmed the environment, whereas 77% of subjects presented with the help vignette thought the chairman did *not* intentionally help the environment.[1]

The asymmetric pattern of responses characteristic of the side-effect effect has been shown to be surprisingly robust. Young, Cushman, Adolphs, Tranel, and Hauser (2006), for example, found that gender, education level, religion, whether English was the primary language, and exposure to moral philosophy had no effect on how subjects respond to the chairman cases. Subjects also exhibit the side-effect effect when aesthetic or prudential (as opposed to moral) badness is at issue (Knobe, 2004a; Knobe & Mendlow, 2004). It appears when Hindi-speaking subjects are given the same vignettes with the word '*jaan-bujhkar*,' the Hindi word for 'intentionally,' substituted for its English equivalent (Knobe & Burra, 2006). Experimental evidence also indicates that ordinary subjects exhibit the side-effect effect as early as four and five years old (Leslie, Knobe, & Cohen 2006).

The robustness of these results led us to suspect that we might be able to find an analogous side-effect effect involving the concept of knowledge. Therefore, we designed an experiment to test the hypothesis that the badness or blameworthiness of a side-effect action undertaken in light of a true belief would have a significant effect on subjects' intuitions about whether that belief counted as knowledge. The participants in our experiment were 749 undergraduate students at [university name omitted for blind review] who received no course credit for their participation. In a classroom setting subjects were randomly assigned either the help or the harm version of the following vignette (adapted from Knobe 2003a):

> The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. We are sure that it will help us increase profits, and 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*. Did the chairman know that the new program would *help/harm* the environment?

Participants were asked to specify their level of agreement regarding the chairman's knowledge using a seven-point Likert scale ranging from -3 to 3, with -3 labeled as "the chairman didn't know," and 3 labeled as "the chairman knew."[2]

Two scoring methods were used to analyze the results. The 'Gradient' scoring method calculated the mean Likert score for subject responses in each experimental condition. The 'All or Nothing' (AON) scoring method grouped together all responses between -3 and -1 as knowledge denials ("the chairman didn't know") and assigned them a new score of 0. Responses ranging from 1 to 3 were interpreted as knowledge attributions ("the chairman knew") and were given a new score of 1. The AON scoring method thus removed participant responses of 0 from the analysis. The mean of the new scores for each experimental condition was then calculated.

On the Gradient scoring method, the mean score for subjects in the help condition was .9060, whereas the mean score for the subjects in the harm condition was 2.2541 (cf. Chart 1 and Table 1). On the AON scoring method, the mean score for subjects in the help condition was .6997, and the mean score for subjects in the harm condition was .9190 (cf. Chart 2 and Table 1). A one-way ANOVA reveals that there is a significant difference between groups: *Gradient* ($F(1, 747)= 102.53$, $P< .0005$) and *AON* ($F(1, 689)= 59.16$, $P< .0005$).

Chart 1 – Gradient

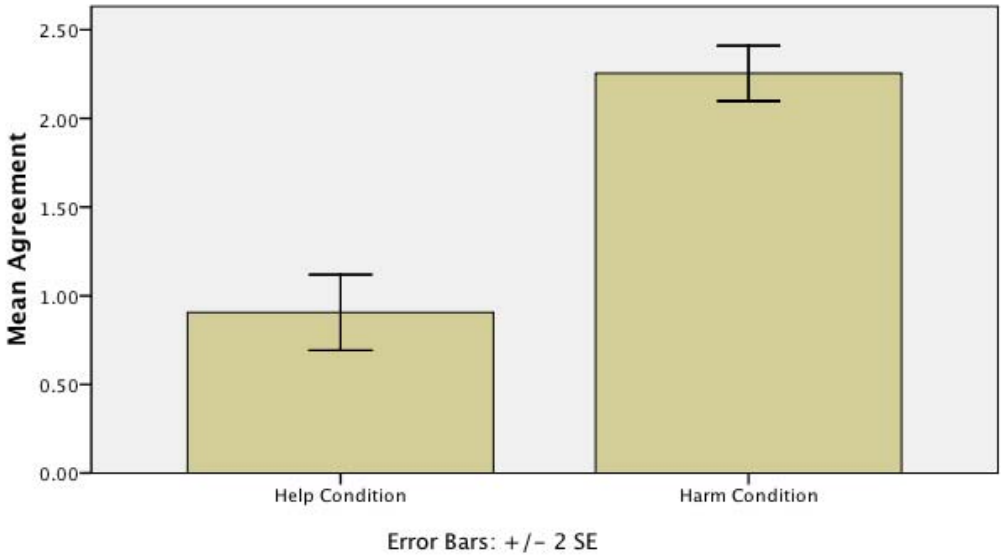
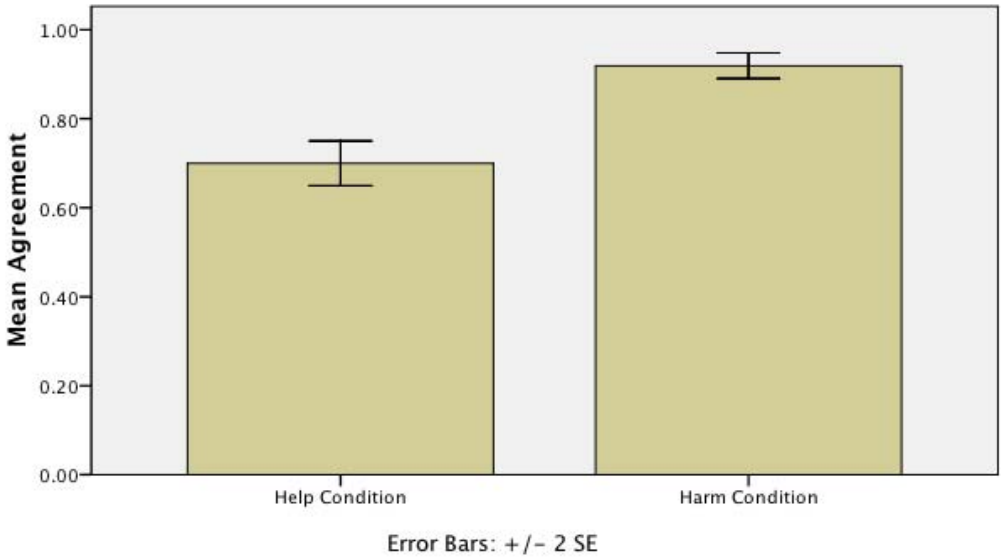


Chart 2 – AON



Survey Type	Scoring	N	Mean	Std. Deviation
HELP	Gradient	383	.9060	2.08499
	AON	333	.6997	.45908
HARM	Gradient	366	2.2541	1.49645
	AON	358	.9190	.27323

Table 1

Like the original side-effect effect results, our findings reveal an asymmetry in subjects' responses. The results are consistent with our hypothesis that subjects are less likely to find that an agent knows an action will bring about a certain side-effect when the effect is good, and more likely to attribute knowledge when the effect is bad. We call the asymmetry in subjects' knowledge attributions 'the epistemic side-effect effect.'

2. Comparison to Other Results and Conjectures in the Literature

Conjectures

Many scholars assume that the chairman knows full well that his actions will have the side-effects described. Hugh McCann (2005, p. 739), for example, claims that "in both vignettes, he knows perfectly well what he is doing." Steve Guglielmo and Bertram Malle (2008) write, "the CEO obviously knows that his action will bring about harm." Jennifer Wright and John Bengson (in press) claim, "In HARM the chairman presumably knew that his action would have a bad outcome."[3] Some scholars take it to be so obvious that the chairman knows the harm will occur

that they mistakenly think the story explicitly ascribes such knowledge to him. Fred Adams (2006, 261), for example, suggests that the reason why most subjects think the chairman intended to harm the environment is "probably because they were told that he knew the harm would occur if they went ahead with the program in question." Careful attention to the original vignettes, however, reveals that while the vice-president testifies to the chairman that the program would harm the environment, no explicit statement of the chairman's knowledge is given.[4] Contrary to these various conjectures and assumptions, our study indicates that subjects do not necessarily attribute knowledge to the chairman and that their epistemic intuitions are influenced by the valence of the side-effect action they consider.

Corroborating Results

In a follow-up study of Knobe's original results, Shaun Nichols and Joseph Ulatowski (2007) gave undergraduates the original chairman of the board cases but also asked them why they answered the prompt questions as they did. They found that most of the subjects who were presented with the harm case and who said the chairman acted intentionally employed epistemic terms to explain their answers. The following answers are typical:

> "He knew the consequences of his actions before he began."
>
> "He knew that implementing the new program would hurt the environment."
>
> "Because he knew he was going to hurt the environment, but chose to do it anyway."
>
> "The chairman knew that this program would harm the environment."
>
> "The chairman intentionally harmed the environment because he was forewarned by the Vice President that the new program would harm the environment."

> "He knowingly made a decision to start the new program after being told that it would harm the environment."

By contrast, epistemic terms are conspicuously absent from the typical answers of those who said the chairman did not act intentionally in the help condition:

> "He didn't care. It was an unintended consequence."
>
> "Because his intention was to make money whether or not it will help the environment."
>
> "He didn't INTEND on helping the environment, he INTENDED on making a profit."
>
> "The chairman's intent was to increase profit."
>
> "His motive was to earn profits."
>
> "His motive was to get as much money as possible."[5]

Our results suggest that this asymmetric use of epistemic terms is not merely characteristic of subjects' post hoc rationalizations of their behavior but is characteristic of their immediate, intuitive responses to the two cases.

In another study, Thomas Nadelhoffer (2006) asked subjects whether the protagonists in the following stories knowingly brought about the deaths of other agents:

> *Case 1:* Imagine that a thief is driving a car full of recently stolen goods. While he is waiting at a red light, a police officer comes up to the window of the car while brandishing a gun. When he sees the officer, the thief speeds off through the intersection. Amazingly, the officer manages to hold on to the side of the car as it speeds off. The thief swerves in a zigzag fashion in the hope of escaping—knowing full well that doing so places the officer in grave danger. But the thief doesn't care; he just wants to get away. Unfortunately for the officer, the thief's attempt to shake him off is successful. As

a result, the officer rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.

Case 2: Imagine that a man is waiting in his car at a red light. Suddenly, a car thief approaches his window while brandishing a gun. When he sees the thief, the driver panics and speeds off through the intersection. Amazingly, the thief manages to hold on to the side of the car as it speeds off. The driver swerves in a zigzag fashion in the hope of escaping—knowing full well that doing so places the thief in grave danger. But the driver doesn't care; he just wants to get away. Unfortunately for the thief, the driver's attempt to shake him off is successful. As a result, the thief rolls into oncoming traffic and sustains fatal injuries. He dies minutes later.

75% of subjects presented with Case 1 said that the thief knowingly brought about the officer's death, whereas only 51% presented with Case 2 said that driver knowingly brought about the car thief's death. These results suggest that subjects' judgments about whether agents knowingly bring about certain results are affected by moral considerations. Although Nadelhoffer tested for attributions of 'knowingly' and we tested for 'didn't know' and 'knew,' both studies indicate that attributions of various epistemic statuses are influenced in surprising ways by the valence of side-effect actions.[6]

Our findings also extend and corroborate recent experimental evidence that the asymmetry characteristic of the side-effect effect can be found in a variety of folk psychological concepts. Tannenbaum, Ditto and Pizarro (2007), for example, presented subjects with the help and harm versions of the chairman of the board vignettes and asked them whether the chairman had a desire to help or harm the environment. On a scale from 1 to 7, the mean response for the help vignette was 1.6, while the mean response for the harm vignette was 3.4. Cushman (2008)

found similar results for 'intend' using 21 different scenarios. Pettit and Knobe (2008) found analogous asymmetries with the concepts of deciding, being in favor of, being opposed to, and advocating. The similarities between these experimental findings strongly suggest that the various side-effect effects stem from general facts about folk psychological attributions rather than from features unique to the individual concepts in question.

Conflicting Results

Knobe and Burra (2006) recently presented the chairman of the board vignettes to Hindi-speaking college students and asked them "Did the chairman *jaan* [the Hindi word for 'know'] that starting the program would harm/help the environment?"[7] 90% of the subjects who were given the help vignette answered "Yes," while 80% of those given the harm vignette also answered "Yes." These results conflict with the asymmetric pattern of results that we found. One factor to consider is that Knobe and Burra's study utilized only 20 subjects, whereas ours employed 749. Further research is needed to determine whether the difference in our results can be chalked up to differences in sample size or whether the epistemic side-effect effect fails to be manifest cross-culturally.

3. Possible Explanations of the Epistemic Side-Effect Effect

Many researchers agree that the most significant differences between the help and the harm conditions of Knobe's original experiment are: (i) the fact that the side-effect in the harm condition is bad but the side-effect in the help condition is good and (ii) the fact that the chairman deserves considerable blame in the harm condition but does not deserve any praise in the help condition. There is, however, widespread disagreement about how to explain the ways

in which these and other factors contribute to the asymmetric pattern found in Knobe's initial side-effect data. Because it seems likely that any successful explanation of the original side-effect effect will be able to shed some light on the epistemic side-effect effect, we turn now to a survey of the leading explanations of Knobe's side-effect effect in the hope of finding explanatory resources that are equal to this task.

Knobe's Original Explanation

Knobe (2003a, 2003b, 2004a, 2004b) originally claimed that the asymmetric results he uncovered were manifestations of folk conceptual competence rather than performance errors. In other words, Knobe claimed that assessments of the badness of actions (and, to a lesser extent, their blameworthiness as well) play a fundamental role in the correct use of the folk concept of intentional action. Knobe's claim ran contrary to the received view that judgments about whether an action was committed intentionally do and should affect our judgments about how much praise or blame to assign but that judgments about praise or blame do not and should not affect our judgments about intentional action.

Most epistemologists would balk at the idea of explaining the epistemic side-effect effect data in an analogous fashion. The suggestion that non-epistemic factors such as the badness or blameworthiness of an action undertaken in light of a belief that *p* can affect whether someone is *properly* taken to know that *p* runs contrary to many deeply held assumptions in epistemology, including the following:

> *Epistemological Purism:* For any two possible subjects *S* and *S'*, if *S* and *S'* are alike with respect to the strength of their epistemic position regarding a true proposition *p*, then *S* and *S'* are alike with respect to being in a position to know that *p*.[8]

Epistemological purists thus maintain that whether a true belief counts as knowledge depends only upon epistemic factors such as evidence or reliability. Purism has been the orthodox view in the epistemological community for centuries. Recently, however, proponents of 'pragmatic encroachment' (e.g., Fantl & McGrath, 2002, 2007; Hawthorne, 2004; Stanley, 2005; Hawthorne & Stanley, in press) have challenged this orthodoxy, arguing that whether a true belief counts as knowledge can depend in part upon non-epistemic facts about how much is at stake for a subject concerning the truth of *p*. It is unlikely, however, that any of these critics would embrace the view that the moral valence of actions performed in light of a belief can affect that belief's status of knowledge. Allowing moral facts to determine the conditions for proper knowledge attribution represents a more significant departure from epistemological purism than does allowing facts about practical rationality to do so. At the same time, however, many defenders of pragmatic encroachment (e.g., Hawthorne and Stanley) place a premium on making their own epistemological theories square with the epistemic intuitions of ordinary subjects. By demonstrating another respect in which folk epistemic intuitions diverge from *a priori* expectations concerning them, our results place pressure on anyone wishing to maintain this combination of views.

Pragmatic Accounts

Adams and Steadman (2004a, 2004b, 2007) argue that the asymmetric responses of the side-effect effect do not indicate anything about the folk's core concept of intentional action. Rather, they claim that the asymmetry is due to pragmatic features of language use and the role these features play in the practice of dispensing praise and blame. According to Adams and Steadman, saying 'You did that on purpose' is a social way of assigning blame and discouraging actions of

which one disapproves. Adams and Steadman consider this role a pragmatic one because blame is not part of the semantic content of 'You did that on purpose' but is nonetheless pragmatically implicated by an assertive utterance of that sentence. They hypothesize that ordinary subjects are inclined to call the chairman's action 'intentional' because denying it was performed intentionally would generate the unwanted implicature that he deserved no blame for what he did. Similarly, they suggest, subjects are disinclined in the help condition to say the chairman performed his action intentionally because saying so would generate the unwanted implicature that he deserved praise for helping the environment. Adams and Steadman contend that because the notion of doing something intentionally or on purpose has this social role, asking people whether an action was performed intentionally may not tap into their core concept of intentional action. It may simply trigger them to think about whether the agents in question should be blamed for their actions.

Someone might offer an analogous explanation of the epistemic side-effect effect by arguing that an assertion of 'You knew that was going to happen' also seems to generate a pragmatic implicature involving blame. In one of the first published critiques of Knobe (2003a), Adams and Steadman (2004a) questioned the limited number of options that subjects were given (only 'intentionally' and 'not intentionally') and hypothesized that if subjects had the option of choosing between 'knowingly' and 'intentionally,' they would choose the former at least as often as the latter. Adams and Steadman suggest that giving subjects the opportunity of saying that the chairman's action was performed knowingly would allow them to express disapproval of his action in the same way that saying he performed it intentionally would. In a somewhat similar vein, Guglielmo and Malle (2008) speculate that the modifier 'knowingly' does not seem to be used for positive intentional actions (e.g., 'He knowingly helped the environment' or 'He

knowingly saved the victim's life'). Rather, it seems to be used for negative intentional actions, especially those involving an omission or an allowing (e.g., 'He knowingly harmed the environment' or 'He knowingly let the victim drown').

In order to test the speculations of Adams and Steadman and Guglielmo and Malle, we performed an archival study using the LEXIS-NEXIS database of general newspaper articles. During the period of July 4th to July 17th, 2008, we found 132 occurrences of the word 'knowingly.' After eliminating 27 duplicate uses, we found that 65 of the remaining 105 uses were associated with (typically criminal) wrongdoing. The agents in question were said to have 'knowingly used false Social Security cards' or 'knowingly allowed a travesty.' These results seem to confirm the suggestions of Adams and Steadman and Guglielmo and Malle that saying that someone knowingly performed an action often performs the social role of assigning blame. However, when we performed a similar study with the word 'knew,' our results were strikingly different. We chose 'knew' rather than 'know' or 'knows' because of the suggestion that a post hoc assertion of 'You knew that was going to happen' may be a way of assigning blame. Of the first 100 non-redundant uses of 'knew' that appeared in publications on July 17th, 2008, only 18 were associated with any wrongdoing. The significant difference between these two sets of findings suggests that there may not be the same connection between negative actions and 'knew' that there is between negative actions and 'knowingly.' If there is not a strong association between negative actions and uses of 'knew,' it will be difficult for those uses to generate conversational implicatures involving blame. Since the epistemic side-effect effect involves 'knew' rather than 'knowingly,' our results thus make it *prima facie* unlikely that someone could successfully explain the effect by appealing to the pragmatic features of language use that Adams and Steadman's claim can explain the basic side-effect effect.[9]

The results of our archival study parallel the experimental results of Knobe (2004b) concerning the disparities that exist between uses of 'intention' and 'intentionally.' Knobe found that, while over 80% of his subjects thought the chairman intentionally harmed the environment, only 29% said it was his intention to do so. And while 20% of those who said the chairman intentionally helped the environment, 0% said it was his intention to do so.[10] Both studies show that one cannot assume *a priori* that words bearing a close etymological relationship will be used by the folk in roughly similar ways.

Distortion Accounts

Some scholars contend that the basic side-effect effect is caused by mechanisms of bias or distortion and thus that subjects' asymmetric responses represent performance errors rather than manifestations of conceptual competence. Malle and Nelson (2003) and Nadelhoffer (2006), for example, claim it is inappropriate for the perceived moral status of an action to have an effect on intentional action attributions in part because they think it is wrong for someone accused of a crime to be more likely to be convicted simply because the crime is immoral. Nadelhoffer (2006, p. 215) claims that in such cases "the cards really are stacked against the defendants from the start." The situation Malle, Nelson and Nadelhoffer describe may well be both legally and morally problematic. However, these normative considerations fail to show that the side-effect effect is a genuine performance error. Perhaps the plain fact of the matter is that the folk concept of intentional action is problematic in precisely these ways. Thus, normative questions about the legal and moral significance of folk concepts should be independent of the purely descriptive question of the actual contours of those concepts.

Following Mark Alicke (2000) and other social psychologists, Nadelhoffer (2006) goes on to hypothesize that folk judgments of blame may result from largely unconscious, spontaneous, affective responses to a harmful event and the people involved. Alicke (2000, pp. 558, 566) writes:

> When a blame-validation mode is engaged, observers review structural linkage evidence 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. In addition to spontaneous evaluation influences, blame-validation processing is facilitated by factors such as the tendencies to over ascribe control to human agency and to confirm unfavorable expectations.
>
> [O]bservers who spontaneously evaluate the actor's behavior unfavorably may exaggerate evidence that established her causal or volitional control and de-emphasize exculpatory evidence.

These reactions also affect observers' judgments by engendering blame-validation processing that subsequently increases the observer's "proclivity to favor blame versus non-blame explanations for harmful events and to de-emphasize mitigating circumstances" (Alicke, 2000, pp. 568-69). Thus, if subjects find an action to be immoral, they will be inclined to seek explanations that favor ascriptions of blame and to overlook explanations that do not. Alicke (2000, p. 557) offers the pessimistic conclusion that "cognitive shortcomings and motivational biases are endemic to blame."[11] Nadelhoffer (2006) argues that since (i) intentional action attributions are influenced by subjects' assignments of blame and (ii) judgments of blame are—at least if Alicke is correct—shot through with distortion and bias, then (iii) the asymmetric

pattern of intentional attributions characteristic of the side-effect effect may be the result of distortion and bias rather than folk conceptual competence.

A proponent of Alicke's affect-driven distortion account might consider arguing that the folk attributions of knowledge characteristic of the epistemic side-effect effect also result from largely unconscious, spontaneous cognitive processes. However, while there may be some plausibility in claiming that the processes underlying judgments of blame involve affective responses to actions and agents, it does not seem at all plausible to claim that the processes underlying knowledge attributions are primarily affective. Yet since affect is what drives Alicke's proposed mechanisms of distortion and bias, the prospects for an analogous explanation of the epistemic side-effect effect do not seem promising.

Someone might argue that even though knowledge attributions are not generally the result of affective responses, such responses may sometimes distort subjects' intuitions about knowledge. Recently, however, a significant neuropsychological challenge to any affect- or emotion-driven distortion account of the side-effect data has been presented by Young et al. (2006). To test the hypothesis that negative emotional responses triggered by morally bad actions affect subjects' tendency to attribute intentionality to agents, Young et al. presented the help and harm versions of the chairman of the board vignettes to subjects with deficits in emotional processing due to damage to the ventromedial prefrontal cortex (VMPC). If normal emotional processing is necessary for the observed asymmetry, individuals with VMPC lesions should show no asymmetry. However, Young et al. found that subjects with VMPC lesions exhibited the same asymmetry in their judgments of intentionality—i.e., they were 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 the

observed asymmetry of intentional action attributions and thus does not mediate the relationship between an action's moral status and its intentional status. To the degree that an affect- or emotion-driven distortion account of the basic side-effect effect data seems unlikely, an analogous account of the epistemic side-effect effect seems unlikely as well.

Furthermore, many epistemologists will object to Alicke and Nadelhoffer's suggestion that any mechanism that lowers evidential standards should count as a mechanism of bias or distortion.[12] Defenders of epistemic contextualism (Cohen, 1988, 1999; DeRose, 1992, 1995; Lewis, 1996), for example, maintain that there are a variety of conversational mechanisms that contribute to the raising and lowering of epistemic standards in contexts of epistemic assessment and that these mechanisms are simply part of our ordinary practices of knowledge attribution. Considering alternate possibilities in which one's belief is false or finding oneself in a high-stakes situation, they claim, has a tendency to raise epistemic standards, while ignoring certain possibilities or remaining in low-stakes situations has a tendency to lower standards. Contextualists contend that which epistemic standards are appropriate in a given context depends upon a variety of factors and that high standards are not always the best or most reasonable ones to employ. Further argument, then, is needed to establish that mechanisms that lower epistemic standards *ipso facto* distort or bias epistemic assessments.

Malle (2006) suggests (but neither endorses nor develops in detail) the following mechanism as a possible mediator of the alleged distortion effects that are representative of the side-effect effect. Perhaps making intentionality judgments and blame judgments side by side while trying to keep them separate taxes subjects' attentional resources. If subjects' decision-making processes are engulfed by the evaluative information presented to them in vignettes, this may lead them to ignore information about the motives, skill and control of agents. If we take a

non-distorted assessment of intentionality to be one that takes into account all available information about motives, skill, control, etc., then to the extent that blame judgments cause subjects to ignore some of this information, they distort assessments of intentionality. A proponent of this hypothesis might consider arguing that knowledge attributions can be distorted by similar factors and that such factors are at least partly responsible for the asymmetric responses that constitute the epistemic side-effect effect. Determining the merits of such a proposal, however, would require more sophisticated experimental methods than the simple survey methods characteristic of recent experimental philosophy.

Malle (2006) also performed an archival study using the LEXIS-NEXIS database and found that 88% of the time the term 'intentional' was used, it was used to refer to a negative action and 99% of the time 'intentionally' was used to refer to a negative action. Malle suggests that strong associations between negative actions and intentionality might lead to biased assessments of intentionality when negative actions are at issue. As noted above, we performed an analogous study which showed a significant disparity between uses of 'knowingly' and 'knew.' Thus, even if Malle were correct in his speculations about the biasing effect of the association between negative actions and 'intentional' and 'intentionally' and even if this account carried over to explain subjects' attributions of 'knowingly,' the archival results we reported cast doubt upon the prospects for an analogous explanation of subjects' use of 'knew.' Either there are no associations between negative actions and 'knew' or the associations are not as manifest as those between negative actions and 'intentionally' or 'knowingly.' In either case, appeals to such associations do not seem to provide a promising avenue to explain the epistemic side-effect effect.

Even though none of the distortion accounts canvassed above seem capable of explaining the epistemic side-effect effect, it is possible that some other distortion account will fare better. In light of the fact that few epistemologists would accept the claim that the asymmetric pattern of responses we have uncovered stems from subjects' competence with the concept of knowledge, the development of such an account would likely be warmly received by the epistemological community. Furthermore, the received wisdom in contemporary epistemology has been that the primary means of explaining away epistemic intuitions that appear problematic for epistemological purism is to argue that the intuitions are responses to pragmatic implicatures rather than to core features of the concepts being deployed (cf. Rysiew, 2001). Psychological research on bias and distortion, however, suggests a wide range of new factors that purists can appeal to in maintaining that certain epistemic intuitions do not reflect deep truths about the concept of knowledge.

Semantic Diversity Accounts

Nichols and Ulatowski (2007) hypothesize that individual differences in how subjects interpret the term 'intentional' may explain both Knobe's original side-effect data and some more recent experimental findings. When Nichols and Ulatowski presented each subject with both the help and the harm versions of the chairman vignette, they found that the overall percentages of subjects who said that the chairman's action was intentional did not change. However, they noticed the following pattern when they analyzed the within-subject data: (i) roughly one third of the respondents said that the chairman's action was not intentional in either the help or the harm conditions, (ii) another third said that his action was intentional in both, and (iii) another third responded asymmetrically. Nichols and Ulatowski were also surprised to find no ordering

effects in their study. Subjects exposed to the help condition first were no less likely to say that the chairman acted intentionally in the harm case, and subjects exposed to the harm condition first were no more likely to say that the chairman acted intentionally in the help condition. Theories that claim that the overall asymmetric pattern of responses to the chairman cases are manifestations of folk conceptual competence (e.g., Knobe, 2003a, 2003b, 2004a, 2004b) typically imply that the responses of those who answer symmetrically should be chalked up to performance error or noise. Nichols and Ulatowski, however, maintain that the systematicity of the symmetric responses should not be ignored—particularly since these responses represent a substantial portion of the data.

Nichols and Ulatowski hypothesize that symmetric responders utilize distinct interpretations of 'intentional'—one (the 'motive' interpretation) that gives pride of place to considerations of an agent's purposes, goals or intentions, and another (the 'foreknowledge' interpretation) that grounds intentionality attributions primarily in agents' justified beliefs about the consequences that will follow from their actions. When asked to explain their responses, subjects who said the chairman did *not* intentionally *harm* the environment focused on his motives, purposes and intentions, offering explanations such as 'He did not set out to harm the environment, he set out to gain a profit' and 'His motivating desire was to make money.' The minority of subjects who agreed that the chairman intentionally *helped* the environment cited his alleged foreknowledge of the side-effects—e.g., 'He knew that the new program would help the environment.'

Nichols and Ulatowski suggest that there are two main ways the foregoing interpretations of 'intentional' can function: (i) Some subjects may generally (or perhaps even always) employ the motive interpretation, while others may generally use the foreknowledge interpretation. (ii)

Some subjects may interpret ‘intentional’ one way in certain contexts (e.g., in the help condition) but interpret it in a different way in other contexts (e.g., in the harm condition). Nichols and Ulatowski claim they know of no evidence that shows one of these types of variation to be more characteristic of the linguistic behavior of ordinary subjects than the other. Consider, however, the hypothesis that the only kind of interpretive variability arises because each subject adopts a single, context-invariant interpretation of ‘intentional.’ While this hypothesis may be able to explain the behavior of those who respond symmetrically, it is decidedly ill-equipped to explain the behavior of asymmetric responders. Thus, it seems that proponents of a semantic variability account should argue that at least some of relevant variability occurs because some of the same subjects interpret ‘intentional’ differently in different contexts.

One problematic feature of Nichols and Ulatowski’s account is that they assume that all research participants will take the chairman to have the same foreknowledge in both the help and the harm conditions. Nichols and Ulatowski then argue that the variation in participant responses is due in part to the fact that only one interpretation of ‘intentional action’ is triggered by these context-invariant facts about the chairman’s foreknowledge. Our research indicates, however, that participants do not uniformly ascribe foreknowledge to the chairman in the two conditions. Nichols and Ulatowski may well be right that distinct conceptions of ‘intentional’ are used by different subjects on different occasions, but the facts about knowledge attribution in the two cases are more complex than their account allows. Furthermore, because Nichols and Ulatowski’s account explains the basic side-effect effect in terms of features that are unique to the concept(s) of intentional action, it does not appear capable of generalizing to the full range of side-effect effects involving the concepts of desire, intending, deciding, being in favor of, being opposed to and advocating.

Fiery Cushman and Al Mele (2008) also claim to have found evidence for more than one interpretation of 'intentional' or, as they put it, for more than one concept of intentional action. Contrary to the traditional view that belief and desire are both necessary conditions for intentional action, Cushman and Mele hypothesize (i) that some subjects treat an agent's desire for a side-effect's occurrence as a necessary condition for that agent's bringing about the side-effect intentionally, (ii) that other subjects treat an agent's justified true belief that a side-effect will occur as a result of the agent's primary action as sufficient for the agent's bringing about that side-effect intentionally, and (iii) that a third group of subjects may treat an agent's justified true belief that a side-effect will occur as sufficient for that agent's bringing about that side-effect intentionally if and only if the side-effect action is morally bad.[13] Cushman and Mele suggest that subjects who thought the chairman intentionally harmed the environment in the harm condition and intentionally helped the environment in the help condition employed the first concept of intentional action and that participants who denied the chairman brought about either side-effect intentionally employed the second. Cushman and Mele tentatively attribute the third concept to those who responded asymmetrically.

Like Nichols and Ulatowski's account, Cushman and Mele's seeks to explain the basic side-effect effect in terms of features that are unique to the concept(s) of intentional action. Thus, it does not appear capable of generalizing to all the other side-effect effects. However, Cushman and Mele's account does seem to have at least some application to the epistemic side-effect effect. Consider the epistemic analogues of the second and third concepts of intentional action proposed by Cushman and Mele. Some subjects may treat an agent's justified true belief that a side-effect will occur as a result of the agent's primary action as sufficient for the agent's knowing the side-effect will be produced. If future research were to reveal that roughly one third

of those given the epistemic versions of the chairman of the board case agreed that he knew in both cases, this proposal might well explain their responses. Other subjects might treat an agent's justified true belief that a side-effect will occur as sufficient for that agent's knowing the side-effect will occur if and only if the side-effect action is morally bad. Ascribing such a concept of knowledge to asymmetric responders in the epistemic side-effect cases might explain their responses as well. Consider now the epistemic analogue of the first concept of intentional action proposed by Cushman and Mele, according to which an agent's desire for a side-effect's occurrence is a necessary condition for knowing that her action would bring about the side-effect. It would be very surprising if a significant portion of ordinary subjects employed such a conception of knowledge, but as the side-effect effect literature has shown, we should never underestimate the potential for folk psychological attributions to surprise us. If further research were to demonstrate the reality of this possibility, the epistemological challenge posed by the epistemic side-effect effect would be significantly deepened.

The leading epistemological account that hypothesizes semantic diversity among ordinary uses of 'knows' is 'epistemic contextualism' (cf. Cohen, 1988, 1999; DeRose, 1992, 1995; Lewis, 1996). Contextualists maintain that 'knows' functions like an indexical, expressing different contents in different contexts. They claim that how strong one's epistemic position with respect to some proposition *p* must be in order to know that *p* varies across contexts. Suppose, for example, that on Friday Keith was planning to stop by the bank to deposit his paycheck. Upon seeing long lines at the bank, Keith drives past the bank and says, "I'll just come back tomorrow. I know the bank will be open on Saturday." If it is not especially important that Keith's check be deposited right away, his utterance will express a different proposition from the one it would express if a great deal depended upon the bank being open on

Saturday—even if Keith's evidence for this belief were identical in the two cases. It is thus possible for an assertion of 'I know the bank will be open on Saturday' to be true in the former context and an assertion of 'I don't know the bank will be open on Saturday' to be true in the latter context without the two assertions contradicting one another. The divergence of meaning of 'know' in the two cases blocks the possibility of contradiction.[14]

While the epistemic side-effect effect data seem to be consistent with the basic claims of epistemic contextualism, extant versions of the view cannot explain the particular pattern of responses that characterize the effect. Contextualists have offered various accounts of the ways in which the content of 'knows' is affected by contexts of use. Cohen (1999, p. 61), for example, writes:

> How precisely do the standards for these predicates get determined in a particular context of ascription? This is a very difficult question to answer. But we can say this much. The standards are determined by some complicated function of speaker intentions, listener expectations, presuppositions of the conversation, salience relations, etc.—by what David Lewis calls the conversational score.
>
> In the case of knowledge ascriptions, salience relations play a central role in determining the standards. In particular, when the chance of error is salient, it can lead knowledge ascribers to intend, expect, presuppose, etc., stricter standards.

In a similar vein, Lewis (1996) formulates eight conversational norms that he claims determine which alternative possibilities may or may not be properly ignored as we go about our ordinary epistemic lives. He maintains that if the sets of permissibly ignored alternative possibilities are different on distinct occasions of epistemic assessment, the contents expressed by 'knows' will also be different. The most important feature of all such contextualist explanations of how the

content of 'knows' changes across contexts is that none of them considers the possibility that the badness or blameworthiness of an action undertaken in light of a belief that *p* might affect whether someone is properly said to know that *p*. In order to explain the epistemic side-effect effect, then, existing versions of contextualism would need to be supplemented by additional explanatory resources. Contextualists have long puzzled over the question of which mechanisms are responsible for lowering contextual standards and have focused instead on the seemingly more tractable problem of specifying the mechanisms responsible for raising epistemic standards. Our results suggest one means by which epistemic standards may be lowered—viz., by discussing bad or blameworthy actions that are connected to the beliefs in question. Thus, the epistemic side-effect effect may provide the basis for further development of the contextualist position.[15]

Knobe's Transgression-Detection Model

Knobe's original explanation of the side-effect effect was that subjects in the harm condition first make a (relatively conscious) judgment that the chairman's side-effect action was bad. This judgment then influences their intuitions about whether the agent in question acted intentionally. More recently, Knobe (2007) claims that his initial hypothesis has been refuted by a series of recent experiments and suggests that the side-effect effect may be due to a nonconscious cognitive process that is dedicated to the detection of norm violations.[16] Knobe hypothesizes that when we consider an action we have an immediate, intuitive reaction that leads to a nonconscious moral judgment about it. These judgments are formed very quickly, taking into account only the first norms that come to mind, and thus involve only a very shallow level of processing. If the action violates the salient norms, it is classified as a transgression, and it is this

judgment that influences subjects' intuitions about intentional action. If subjects take the time to think more deeply about the behavior, they may reflect on a variety of additional considerations and come to a conscious, considered judgment about the moral status of the behavior that may conflict with the nonconscious judgment. Knobe further hypothesizes that the nonconscious judgment may remain in memory even after the conscious judgment has been made and may continue to affect subjects' intuitions about whether certain actions are intentional.

The transgression-detection model allows Knobe to explain subjects' responses to cases in which an agent's action is clearly good and yet the side-effect effect is found. For example, Knobe (2007) gave subjects either the violation or the fulfillment version of the following vignette:

> 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 *violating/fulfilling* the requirements of the racial identification law." The CEO said: "Look, I know that I'll be *violating/fulfilling* the requirements of the law, but 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.

Subjects were asked whether the CEO acted intentionally. 81% of the subjects said he violated the law intentionally, but only 30% said he fulfilled the law intentionally. These results conflict with Knobe's previous explanation of the side-effect effect, which predicts that bad actions are

more likely to be categorized as intentional than good ones. In the present case, violating the racial identification law is good, yet subjects are more likely to judge the violation to be intentional. Knobe's new hypothesis can explain this data, however, because the CEO has violated a salient norm.

Knobe (2007, pp. 102-103) maintains that his theory of transgression detection is not meant to be relevant only to questions about intentional action intuitions. If it is correct, he suggests that "what we are uncovering here is a general truth about how people make moral judgments." Further research is needed to determine whether the epistemic side-effect effect always manifests itself when norm violations are present. If such a correlation could be established, it would lend significant support to Knobe's speculation about uncovering a general and far-reaching truth about moral cognition.

Summary

Of the many explanations of the original side-effect effect that have been offered, most seem to have some potential to be developed into (at least partial) explanations of the epistemic side-effect effect.[17] The use of some of these explanations in this capacity, however, such as Knobe's original conceptual competence proposal, would require significantly rethinking the concept of knowledge. Epistemological purists (and, indeed, many of their critics) will most likely favor the development of a new kind of distortion account that limits the range of non-epistemic factors that can properly influence a true belief's status as knowledge. While semantic diversity accounts fit comfortably with epistemic contextualism, many contextualists may be uncomfortable with the degree to which such a marriage allows the 'pragmatic encroachment' of

non-epistemic factors into the set of knowledge-determining conditions. Pragmatic accounts seem to represent the least promising category of explanation.

We favor explanations of the epistemic side-effect effect that, like Knobe's transgression-detection account, hypothesize that the various side-effect effects (e.g., those involving the concepts of intentional action, desire, intending, deciding, being in favor of, being opposed to, advocating and knowing) arise from general facts about the relationship between folk psychology and normative assessment rather than from features that are unique to the individual concepts in question. Pettit and Knobe (2008) suggest, "Not only does the impact of moral judgment extend beyond the concept of intentional action, moral judgments appear to be having some impact on just about every concept that involves holding or displaying a positive attitude toward an outcome." We endorse a generalized version of this thesis that looks to the full range of normative judgments (both moral and non-moral) that ordinary subjects make as factors that can influence the attribution of pro-attitudes. Not only do our findings seem to provide some degree of confirmation for this thesis, we hope they will also generate increased interest in such a thesis as well.

4. Conclusions and Future Research

Consider the following, widely endorsed theses: (i) Whether a true belief counts as knowledge depends only upon epistemic factors such as evidence or reliability. (ii) Because the target of philosophical analyses of knowledge is the ordinary person's concept of knowledge, such analyses should be answerable to data about "what the ordinary person would say" in response to various epistemological thought experiments. Our work contributes to a growing body of research that makes the conjunction of (i) and (ii) increasingly difficult to maintain. Weinberg,

Stich and Nichols (2001; Nichols, Stich, & Weinberg, 2003), for example, found that epistemic intuitions vary with the cultural background, education background and socioeconomic status. While it may be possible to dismiss a small class of such findings as due to performance errors or noise, as more and more experimental data is gathered that shows that ordinary subjects' knowledge attributions are influenced by a variety of non-epistemic factors, this line becomes ever more difficult to maintain. A more promising strategy may be to reject the idea that an account of how we should think about knowledge must answer to the epistemic intuitions of ordinary subjects.

Our research also suggests that the practice of making epistemic evaluations may be more closely related to our practices of assigning moral praise or blame than many epistemologists have thought. According to the traditional view of the relationship between knowledge and action, whether a subject, *S*, knows that *p* is completely independent of whatever actions *S* may undertake in light of *S*'s belief that *p*. Recent proponents of 'pragmatic encroachment' in epistemology (e.g. Fantl & McGrath, 2002, 2007; Hawthorne, 2004; Stanley, 2005), however, have challenged this view. Stanley (2005), for example, argues that if the warranted expected utilities of the actions at a subject's disposal are significantly affected by whether or not a proposition is true, then that subject must satisfy higher epistemic standards in order to know that proposition than another subject whose warranted expected utilities are not significantly affected. In short, "the greater the practical investment one has in a belief, the stronger one's evidence must be in order to know it" (Stanley, 2005, p. 88).[18]

In opposition to the traditional view that connects action with subjective credence but not with knowledge, proponents of pragmatic encroachment have also defended various knowledge-action principles. Fantl and McGrath (2007, p. 559), for example, propose the following:

(KA) *S* knows that *p* only if *S* is rational to act as if *p*.

Stanley (2005, p. 9) contends:

> [I]t is immensely plausible to take knowledge to be constitutively connected to action, in the sense that *one should act only on what one knows*.

Hawthorne (2004, p. 30) suggests that knowledge is connected to practical deliberation in the following manner:

> On the face of it, then, we operate with a conception of deliberation according to which, if the question whether *p* is practically relevant, it is acceptable to use the premise that *p* in one's deliberations if one knows it and (at least in very many cases) unacceptable to use the premise that *p* in one's practical reasoning if one doesn't know it. At a rough first pass: one ought only to use that which one knows as a premise in one's deliberations.

More recently, Hawthorne and Stanley (in press) defend the following Action-Knowledge Principle:

(AKP) Treat the proposition that *p* as a reason for acting only if you know that *p*.

Defenders of pragmatic encroachment have thus sought to identify various non-traditional ways in which the truth value of ordinary knowledge attributions are properly affected by various factors concerning the actions and practical deliberations of potential knowers. We suggest that the epistemic side-effect effect may reveal a further respect in which knowledge is connected to action: whether a subject knows that *p* may depend upon the moral status of actions the subject performs in light of the belief that *p*.

In future research we plan to investigate whether there are other features of actions and actors that can affect knowledge attributions, such as the importance of an action, the degree to which a protagonist identifies with the action, and the protagonist's perceived moral character.

In addition, we intend to test the hypothesis that an agent's desire for a side-effect's occurrence can influence whether that agent is taken to know the side-effect will occur. In order to test the applicability of Knobe's transgression-detection model to the epistemic side-effect effect, we will investigate whether or not the violation of a salient social norm affects subjects' attributions of knowledge. We also plan to explore the possibility that similar side-effect effects can be demonstrated in cases that do not involve folk psychological notions at all. For example, we speculate that folk attributions of causality may also be sensitive to the valence of the actions or events in question. We believe that the various side-effect effects that have been demonstrated in the literature point to general features of human psychology that may require us to rethink significant portions of our conception of ourselves.

References

Adams, F. (2006). Intentions confer intentionality upon actions: A reply to Knobe and Burra. *Journal of Cognition and Culture*, 6, 132-146.

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.

Adams, F., & Steadman, A. (2007). Folk concepts, surveys, and intentional action. In C. Lumer (Ed.), *Intentionality, deliberation, and autonomy: The action-theoretic basis of practical philosophy* (pp. 17-33). Aldershot: Ashgate.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, 126, 556-574.

Alicke, M. D. (in press). Blaming badly. *Journal of Cognition and Culture*.

Buckwalter, W. (2008). Knowledge isn't closed on Saturdays. Unpublished manuscript.

Cohen, S. (1988). How to be a fallibilist. *Philosophical Perspectives*, 2, 91-123.

Cohen, S. (1999). Contextualism, skepticism, and the structure of reasons. *Philosophical Perspectives*, 13, 57-89.

Cushman, F. (2008). The effect of moral judgment on causal and intentional attribution: What we say, or how we think? Unpublished manuscript.

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

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

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

Doris, J., Knobe, J., & Woolfolk, R. L. (2007). Variantism about responsibility. *Philosophical Perspectives*, 21, 183-214.

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.

Guglielmo, S., & Malle, B. F. (2008). Can unintended side-effects be intentional? Solving a puzzle in people's judgments of intentionality and morality. Unpublished manuscript.

Hawthorne, J. (2004). *Knowledge and lotteries*. New York: Oxford.

Hawthorne, J., & Stanley, J. (in press). Knowledge and action. *Journal of Philosophy*.

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. (2004a). Folk psychology and folk morality: Response to critics. *Journal of Theoretical and Philosophical Psychology*, 24, 270-279.

Knobe, J. (2004b). Intention, intentional action and moral considerations. *Analysis*, 64, 181-187.

Knobe, J. (2007). Reason explanation in folk psychology. *Midwest Studies in Philosophy*, 31, 90-107.

Knobe, J., & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition*, 6, 113-132.

Knobe, J., & Mendlow, G. (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.

Leslie, A., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment. *Psychological Science*, 17, 421-427.

Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74, 549-567.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language*, 23, 165-189.

Malle, B. F. (2006). Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture*, 6, 87-113.

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, 563-580.

Mallon, R. (in press). Knobe vs. Machery: Testing the trade-off hypothesis. *Mind and Language*.

McCann, H. (2005). Intentional action and intending: Recent empirical studies." *Philosophical Psychology*, 18, 737-748.

Nadelhoffer, T. (2004a). Praise, side effects, and intentional action. *Journal of Theoretical and Philosophical Psychology*, 24, 196-213.

Nadelhoffer, T. (2004b). Blame, badness, and intentional action: A reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology*, 24, 259-269.

Nadelhoffer, T. (2005). Skill, luck, and folk ascriptions of intentional action. *Philosophical Psychology*, 18, 343-354.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality. *Philosophical Explorations*, 9, 203-220.

Nichols, S., Stich, S., & Weinberg, J. (2003). Metaskepticism: Meditations in ethno-epistemology. In S. Luper (Ed.), *The Skeptics: Contemporary Essays* (pp. 227-247). Burlington,VT:Ashgate Press.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind and Language*, 22, 346-365.

Pettit, D., & Knobe, J. (2008). The pervasive impact of moral judgment. Unpublished manuscript.

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, 291-298.

Pizarro, D., Knobe, J., & Bloom, P. (2008). College students implicitly judge interracial sex and gay sex to be morally wrong. Unpublished manuscript.

Rysiew, P. (2001). The context-sensitivity of knowledge attributions. *Noûs*, 35, 477-514.

Sinnott-Armstrong, W., Mallon, R., McCoy, T., & Hull, J. (in press). Intention, temporal order, and moral judgments. *Mind and Language*.

Stanley, J. (2005). *Knowledge and practical interests*. New York: Oxford University Press

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

Weinberg, S., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29, 429-460.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, 100, 283-301.

Wright, J., & Bengson, J. (in press). Asymmetries in judgments of responsibility and intentional action. *Mind and Language*.

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.

[*] Co-authorship is equal.

[1] Knobe (2003a) also found that there was a significant correlation between the degree to which subjects attached praise or blame to an agent's actions and their judgments about whether the side-effects were brought about intentionally.

[2] Other recent studies in experimental epistemology (e.g., Weinberg, Nichols, & Stich, 2001; Nichols, Stich, & Weinberg, 2003) give subjects a forced-choice between saying that a protagonist 'really knows' or 'only believes' that something is the case. We think this methodology is potential problematic because subjects may interpret 'really knows' to mean 'knows with certainty' or at least 'knows' according to some higher standard than normal.

[3] Cushman & Mele (2008, 172) write, "One [CEO] knowingly *harms* the environment by starting a profit-making venture and does not "care at all" about harming it; the other knowingly *helps* the environment by starting a profit-making venture and does not "care at all" about helping it."

[4] In the second experiment Knobe (2003a) originally used to demonstrate the side-effect effect includes the central protagonist self-attributing knowledge:

> A lieutenant was talking with a sergeant. The lieutenant gave the order: "Send your squad to the top of Thompson Hill." The sergeant said: "But if I send my the squad to the top of Thompson Hill, we'll be moving the men directly *into/out of* the enemy's line of fire. Some of them will surely be *killed/rescued*!" The lieutenant answered: "Look, I know that they'll be *in/taken out of* the line of fire, and I know that some of them will be *killed/otherwise*. But I don't care at all about what happens to our soldiers. All I care about is taking control of Thompson Hill." The squad was sent to the top of Thompson Hill. As expected, the soldiers were *moved into/taken out of* the enemy's line of fire, and some of them were *killed/thereby escaped getting killed*.

Because of the structural similarity of the corporate and military cases, one might have expected that subjects would automatically assume the chairman knew his company's new program would have the stated consequences. Our results show this is not the case.

[5] Similar results are reported in Mallon (in press).

[6] Cf. the results of the archival study we report below for an important difference between uses of 'knew' and 'knowingly.'

[7] An odd feature of Knobe and Burra's study is that their vignettes were presently entirely in English and the questions they gave to subjects were asked were almost entirely in English—the only exception being the single Hindi word '*jaan*' in the middle of the sentence.

[8] Cf. Fantl and McGrath (2007, p. 558). Stanley (2005) calls the same view 'intellectualism.'

[9] A recent study by Guglielmo and Malle (2008, p. 25ff.) reveals that one must be careful not to assume there exists too close a relationship between uses of 'intentionally' and 'knowingly.' Guglielmo and Malle gave subjects the original chairman prompts but included the following additional dialogue between the CEO (i.e., the chairman) and the vice-president in the harm condition:

> After the vice-president informs the CEO that the new program will help us increase profits but also harm the environment, the CEO asks, "Are there any other options?" The vice-president replies, "I have looked at the alternatives, and there is one older program that would increase profits slightly less but not harm the environment as much."

In one experimental condition the CEO rejects this alternative: "Well, I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program." In another condition he accepts the alternative: "Well, I don't care at all about harming the environment. I just want to make as much profit as I can. But that older program sounds fine too. Let's start that program." In addition to being asked standard questions about whether the CEO intentionally harmed the environment and whether he deserves blame for his actions, subjects were also given multiple action descriptions and asked to select the "most accurate description of what the CEO did" from the following options:

> "The CEO willingly harmed the environment."
>
> "The CEO knowingly harmed the environment."
>
> "The CEO intentionally harmed the environment."
>
> "The CEO purposefully harmed the environment."

Across conditions Guglielmo and Malle found that only 1% of respondents selected "The CEO intentionally harmed the environment" as the most accurate description. This result is quite surprising in light of Knobe's results. By contrast, 82% found "The CEO knowingly harmed the environment" to be the most accurate, and 14% opted for the description involving "willingly." Guglielmo and Malle (2008, pp. 27-28) write, "This finding is particularly noteworthy because the same participants who first provided the forced choice intentionality response afterwards picked the most accurate behavior description. Not even priming or pressures of consistency could convince people to describe the protagonist as having intentionally harmed the environment, if a different description was available."

[10] Cf. McCann (2005) for similar results.

[11] Alicke (in press) also claims that judgments about intentional action are also susceptible to 'outcome bias'—viz., the susceptibility of post hoc judgments of intentionality to be influenced by knowledge of how an action actually turned out. Such judgments can be influenced by factors independent of the agent's decision-making process and intentions. Alicke contends that when subjects in Knobe's original studies were asked whether the chairman *intended* or had the *intention* to harm the environment, they are being asked a question about his thought process and desires prior to the action being undertaken. However, when they are asked whether he harmed the environment *intentionally*, they are answering a question that includes the outcome that he did harm the environment.

[12] Malle (2006) also suggests that evidential standards for intentionality attributions are lowered when negative actions are being considered.

[13] Cushman and Mele are not equally confident about this third possibility. Hence, the 'two-and-a-half folk concepts' in the subtitle of their article.

[14] Contextualist proposals have been made in other areas of experimental philosophy as well. Woolfolk, Doris and Darley (2006, p. 298), for example, suggest that "differing considerations are salient to moral responsibility attribution in different contexts, and that patterns of responsibility attribution may also vary culturally and developmentally." Doris, Knobe and Woolfolk (2007) defend a version of this view, which they call 'variantism about responsibility.'

[15] But cf. Buckwalter (2008) for an important experimental challenge to epistemic contextualism.

[16] Cf. Nichols & Ulatowski (2007), Machery (2008), Phelan & Sarkissian (2008), Sinnott-Armstrong, et al. (in press), Wright and Bengson (in press), Pizarro, et al. (2008), and Cushman (2008).

[17] Some prominent accounts that we have not discussed include Machery (2008), Wright & Bengson (in press) and Sinnott-Armstrong et al. (in press). While we have only reflected on the possibility of extending extant explanations of the basic side-effect effect to the epistemic side-effect effect, other kinds of explanations are clearly possible.

[18] Cf. Buckwalter (2008) for a recent challenge to the claim that raising the practical stakes leads ordinary subjects to raise the standards for knowledge.