

The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it

Mark T. Phelan · Hagop Sarkissian

Received: 9 July 2006 / Accepted: 4 September 2006
© Springer Science+Business Media B.V. 2006

Abstract Recent and puzzling experimental results suggest that people's judgments as to whether or not an action was performed intentionally are sensitive to moral considerations. In this paper, we outline these results and evaluate two accounts which purport to explain them. We then describe a recent experiment that allegedly vindicates one of these accounts and present our own findings to show that it fails to do so. Finally, we present additional data suggesting no such vindication could be in the offing and that, in fact, both accounts fail to explain the initial, puzzling results they were purported to explain.

Keywords Experimental philosophy · Action theory

I

How does a person's evaluation of a particular effect influence her judgment about whether it was brought about intentionally? Until fairly recently, prominent theories maintained that intentionality judgments hinge on assumptions about the causal relationship between agents and effects (e.g., Adams, 1986, 1997; Davidson, 1963; McCann, 1986). For example, an effect would be judged intentional only when the agent foresees that the effect will result from her actions and is trying to bring it about. According to such theories, evaluations concerning the goodness or badness of effects or agents are strictly irrelevant to judgments of intentionality. Knobe (2003) showed this not to be the case. Knobe asked subjects to reflect on the fol-

M. T. Phelan (✉)
Department of Philosophy, University of North Carolina, Caldwell Hall, Chapel Hill, NC
27599, USA
e-mail: mphelan@email.unc.edu

H. Sarkissian
Department of Philosophy, Duke University, West Duke Building, 27708, Durham, NC, USA
e-mail: hss12@duke.edu

lowing vignette (or one exactly similar to it, except that the word “harm” was replaced with the word “help” throughout):

The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.’

The chairman of the board answered, ‘I don’t care at all about harming the environment. I just want to make as much profit as I can. Let’s start the new program.’ They started the new program. Sure enough, the environment was harmed.

Subjects were then asked whether or not the chairman intentionally harmed (or helped) the environment. The vast majority of subjects presented with the harm condition judged that the chairman intentionally harmed the environment, whereas very few judged that the chairman intentionally helped the environment.

Harming/helping the environment in this case has been called a *side effect*—an effect that (a) an agent does not try to bring about, and (b) does not constitute a means to some effect she is trying to bring about. Think of the bystander throwing the switch in the trolley car cases: The bystander is not trying to cause the death of the one innocent person on the spur track she sends the trolley down; it’s simply that his death results from (though is not a means to) saving the five people on the track from which the train has just been diverted. Similarly, it seems clear in the case above that the chairman was not trying or desiring to bring about the effect of harming (or helping) the environment. (In fact, he says this more or less directly in the vignette.) Therefore, according to the traditional criteria for intentionality judgments, there should be no difference between subjects’ judgments concerning intentionality in the two cases. In other words, the perceived value of the effect in question should be irrelevant. However the prediction of the traditional criteria did not pan out. The results show that people’s evaluative attitudes do influence their judgments of intentionality.

Yet what is the object of these evaluations? There are many possible candidates. For example, people might be judging that the act itself was wrong; or that the consequences were bad; or that the chairman had a nasty character; or that the act transgresses communal expectations; or any combination of these. As it stands, the literature has focused on two possible accounts.

Knobe favors an account according to which, out of all the possible objects of moral evaluation, subjects are cuing in on the badness of the side effect itself. His basic thesis holds that if subjects apprehend a side effect as bad, they will then focus on whether it was foreseen. If so, the side effect is judged to have been brought about intentionally (Knobe, 2003; 2005; Knobe & Mendlow, 2004). Because it focuses on the badness of side effects in order to explain intentionality judgments, let’s call Knobe’s model the badness model of intentional side effects, or BAM, for short.¹

A competing account focuses on the blameworthiness of the agents that produce the side effects. According to the blameworthiness model, or BLAM, if subjects

¹ Because the harming-chairman case, the one for which judgments of intentionality increased, involves what is most naturally thought of as a morally bad side effect, one might conclude that Knobe intended *moral* badness as inclining subjects to judgments of intentionality. Indeed, Knobe (2003) is most naturally read as a defense of this thesis. However, Knobe in fact holds a weaker thesis, such that some sense of badness (and not simply moral badness) can lead to intentionality judgments (Knobe & Mendlow, 2004).

apprehend a side effect as bad, they will consider the agent's stance or attitude toward that side effect. If this attitude signals a good or bad character—in other words, if the person is praiseworthy or blameworthy—the side effect is judged to have been brought about intentionally. So people praise or blame the agent before they settle the question of intentionality and, having already decided that people deserve praise or blame, BLAM suggests that subjects are, in an exercise of post-hoc justification, inclined to judge that the agent acted intentionally.² We class Nadelhoffer (2004a–c) among the proponents of this view.

BAM and BLAM, the competing theories of what leads to judgments of intentionality, are not merely technical theories about the (somewhat scholarly) question of what features people exploit in their intentionality judgments. Which theory turns out to be correct also bears on the normative question: To which features *should* our concept of intentionality be sensitive. The relevance of the technical debate to the normative question rests on the following argument. It is reasonable to suppose, *prima facie*, that a concept should be sensitive to those features which it is actually sensitive to in its central uses. One of the central uses of the concept of intentional action is in assessments of praise and blame. According to the BAM model, the evaluative judgments that influence ascriptions of intentionality can figure in assessments of blame or praise. Therefore, according to BAM, evaluative judgments influence ascriptions of intentionality in cases where the concept of intentionality is being put to one of its central uses. In other words, if BAM is correct, and the argument just canvassed holds, evaluative judgments in cases such as the chairman cases *should* influence judgments of intentionality. Thus, BAM imitates previous positions in holding that the concept of intentional action bears important and fundamental connections to moral concepts (Harman, 1976; Lowe, 1978; Pitcher, 1970). On the other hand, BLAM cannot hold that the evaluative judgments it takes to influence ascriptions of intentionality do so when the concept is put to the central use of *assessing* praise or blame. The evaluative judgments BLAM takes to be relevant to ascriptions of intentionality *just are* judgments of praise or blame. To hold that intentionality judgments so influenced go on to influence assessments of praise or blame would be circular.³ More to the point, the intentionality judgments BLAM focuses on are the result of post-hoc reasoning, which suggests that they are

² Recently, the psychologist Jonathan Haidt has outlined a 'social intuitionist' model of moral judgment that affords a more general framework within which BLAM might be situated (Haidt, 2001). On this model, the vast majority of people's moral judgments are driven by initial flashes of intuition (what we might call 'gut feelings') that do not result from any process of reasoning. People can, of course, provide reasons to support their moral judgments when prompted to do so, but all such reasoning is post-hoc justification of a prior moral judgment. Only in rare cases does reasoning play a constitutive role in moral judgments. Similarly, for BLAM, subjects first have a strong inclination to blame the agent who engendered the bad side-effect in question, and this prior moral judgment then leads them to say—post-hoc—that the agent brought about the side-effect intentionally.

³ According to BLAM, judgments of blameworthiness precede judgments of intentionality in the relevant side effect cases. According to BAM, they do not. An anonymous reviewer for *Philosophical Studies* suggests the following method for generating evidence about who is right: "Show two groups the same story. Instruct them to answer as soon as they can. (They click a box on a computer monitor. Their clicks are timed, as is appearance of the story on the monitor.) Ask one group whether the agent A-ed intentionally, and ask the other whether he should be blamed." This is an intriguing method and presents an interesting proposal for future work. If it turned out that people regularly made judgments of one kind before the other, this might endorse one view over the other.

not among the central uses of the concept. BLAM is thus naturally associated with views according to which moral considerations lead subjects to disregard those considerations which are appropriate to judgments of intentionality, such as foresight and trying (c.f. Mele & Sverdlik, 1996). These views maintain that examples such as the chairman cases do not reveal connections between the concept of ‘intentional action’ and moral concepts. Instead, they represent misapplications of the concept.⁴

II

So much for our brief explication of BAM and BLAM. How might one adjudicate between them? If one could find test cases in which subjects judge a particular side effect to be both intentional and bad, but do not judge the agent to be blameworthy (or vice-versa), one would have shown—at least *prima facie*—one theory and not the other to be correct. Such a test case is suggested by Knobe and Mendlow (2004). They asked 20 subjects in a New York park about the following vignette:

Susan is the president of a major computer corporation. One day, her assistant comes to her and says, “We are thinking of implementing a new program. If we actually do implement it, we will be increasing sales in Massachusetts but decreasing sales in New Jersey.”

Susan thinks, “According to my calculations, the losses we sustain in New Jersey should be a little bit smaller than the gains we make in Massachusetts. I guess the best course of action would be to approve the program.”

“All right,” she says. “Let’s implement the program. So we’ll be increasing sales in Massachusetts and decreasing sales in New Jersey.” (Knobe & Mendlow, 2004)

Of those tested, 80% thought Susan deserved neither praise nor blame for her action, while 75% thought she decreased sales intentionally. Why did subjects judge the side effect to be intentional in the absence of judgments of praise or blame? Knobe and Mendlow claim that there is a natural sense in which ‘decreasing sales in New Jersey’ might be thought of as bad, that subjects were picking up on this badness and, as BAM predicts, this evaluation of the side effect influenced their judgments of intentionality, even in the absence of blameworthiness. Thus, they conclude that BAM is vindicated.

Knobe and Mendlow’s claim that subjects are picking up on a sense of badness is essential to their argument. Since this is a claim about folk psychology, we thought the best way of testing it would be to ask the folk. To that end, we reran the test.

⁴ Nadelhoffer (2004c) suggests a role for the post-hoc intentionality judgments his theory posits: “the concept of intentional action...could still be used to amplify, verify, mitigate, or exculpate our antecedent attributions of moral responsibility. And even though in these situations our notion of intentional action would admittedly not play its usual role of fixing blame, it would nevertheless have an important role to play” (262). Important, perhaps, but arguably not one of its central roles. In any case, our discussion was only meant to point out the affinity BAM and BLAM have to various theories of the conceptual role of “intentionality.” It is not essential that proponents of either BAM or BLAM accept the one account or the other.

However, instead of asking only whether Susan deserved any blame and if she acted intentionally, we also asked subjects whether decreasing sales in New Jersey was bad.⁵ Our subjects were 36 students in several introductory philosophy classes at the University of North Carolina, Chapel Hill. Nearly all of them (94.5%) responded that Susan deserved neither praise nor blame for decreasing sales (the remaining 5.5% thought she deserved praise), and a clear majority (64%) responded that she decreased sales intentionally. This is more or less in line with Knobe and Mendlow's original findings. However, and most interestingly, a mere 14% responded that decreasing sales was bad. Thus, whatever explains the high proportion of intentionality judgments in this case, it is presumably not the perceived badness of the side effect.

A BAM theorist might attempt to explain our data by alleging that, despite the specificity of our question as to whether or not "decreasing sales in New Jersey" was bad, subjects were in the grip of a holistic consideration of the case when they answered. A BAM theorist might contend that decreasing sales in New Jersey *is* bad, subjects *do* apprehend it as such, and their appreciation of the badness of decreasing sales *does* lead them to judge that Susan acted intentionally. However, when asked whether decreasing sales was bad, they mistakenly think of the overall increase of sales, and so they do not answer that the decrease was bad. Another possibility is that the phrasing "bad" in this question is too vague. Perhaps we need something that more naturally picks out the purported natural sense in which decreasing sales is bad. Maybe we should have asked subjects if, considered on its own, it would be bad to decrease sales in New Jersey.

One way of responding to both of these suggestions would be to formulate a case that only differed from Knobe and Mendlow's decreasing sales case in that there is no natural sense in which the side effect could be considered bad. We could then argue by analogy that the judgment of intentionality made regarding the case is made on the same basis as the judgment in its duplicate. If there were no sense in which the side effect were bad, and people did not judge it to be bad, then we would be justified in concluding that considerations of badness play no part in leading people to judge the side effect in either case intentional. Now, there is arguably a natural sense in which decreasing sales is bad, but there is surely no *natural* sense in which it is bad to increase the prominence of one division in a company compared to another division in the same company. Therefore we asked 33 people spending time on Duke University campus about that effect in the following case:

Susan is the president of a major computer corporation. One day, her assistant comes to her and says, "We are thinking of implementing a new program. If we actually do implement it, we will be increasing sales in our Atlantic division, but this will also increase the prominence of the Atlantic division compared to the Pacific division."

Susan thinks, "According to my calculations, increasing the prominence of the Atlantic division compared to the Pacific division should not affect our other programs, and we will be making gains in the Atlantic division. I guess the best course of action would be to approve the program."

⁵ This strategy is suggested, but not implemented, in Nadelhoffer (2004c).

“All right,” she says. “Let’s implement the program. So we’ll be increasing sales in our Atlantic division, but this will also increase the prominence of the Atlantic division compared to the Pacific division.”

In this case the relevant effect is to increase the prominence of the Atlantic division relative to the Pacific. Now, it does not seem that Susan is trying to bring about the relevant effect in this case. After all, Susan considers the effect in this case in the same way as she considers the effect in the previous case. Also, it does not seem reasonable to suppose that people would think of increasing the prominence of one division relative to another as bad. Indeed, subjects’ responses bore out this last assumption. Hardly anyone (only 6%) thought “increasing the prominence of the Atlantic division relative to the Pacific division” was bad, and hardly anyone (only 3%) found it blameworthy.⁶ Nevertheless, similar to the previous case, approximately 67% responded that Susan acted intentionally.

In both cases (the original and our variant) we see the same pattern: hardly any subjects think the effects are bad, hardly any hold the agents blameworthy, but they nonetheless judge the effects to have been done intentionally. It cannot be recognition of the badness of the side-effect which explains people’s judgments of intentionality in the increasing prominence case. By analogy, it is not recognition of the badness of the side-effect which explains people’s judgments of intentionality in Knobe and Mendlow’s original decreasing sales case. Thus, Knobe and Mendlow’s case does not constitute evidence in favor of BAM. Indeed, the case is strictly irrelevant to the BAM/BLAM debate. Whatever explains judgments of intentionality in that case, it is some non-moral feature.⁷

III

We now turn away from the question of whether Knobe and Mendlow’s case actually vindicates BAM, and instead present a direct challenge to that theory. We asked 21 subjects about the following vignette:

The city planner’s assistant came to him and said, “We finally developed a plan to address our pollution problem. We have a new reclamation project that will clean up the toxic waste polluting the former industrial area. However, if we actually do implement the project, it will also increase the levels of joblessness.”

⁶ In fact, 42% of subjects judged the side effect in the increasing prominence case to be good, and 40% thought Susan deserved praise for it. We were a little surprised by these numbers. Could these judgments be affecting ascriptions of intentionality? This cannot be ruled out. However, even if this were true, neither BAM nor BLAM could account for it. According to BAM, if people judge an effect to be good, then they must examine whether the agent was trying to bring it about (and whether or not it was a means to some other end the agent was trying to bring about) in order to assess whether it was brought about intentionally. But it’s clear that Susan is not trying to bring about the increase in prominence, so BAM cannot account for this result. According to BLAM, an agent’s pro attitude toward a good effect can lead to ascriptions of intentional action (Nadelhoffer, 2004b). However, Susan displays no such pro-attitude towards the effect. So BLAM cannot account for this result either.

⁷ Nadelhoffer (2006) ran a study with a similar structure and obtained a similar result. The study concerned a sniper who fired his rifle and, as a side effect, ended up heating the barrel of his gun. Of those asked, 68% deemed the sniper to have heated his barrel intentionally. This represents another instance of a neutral side effect that is nonetheless deemed intentional. Our thanks to an anonymous reviewer for *Philosophical Studies* for pointing this out.

The city planner answered, “I feel terrible about increasing joblessness. But we have to do something about our pollution problem. Let’s start the project.”

They started the project. Sure enough, the toxic waste polluting the former industrial area was cleaned up, and joblessness levels increased.

Out of those asked, 67% answered that, yes, it was bad that the city planner had increased levels of joblessness. If BAM were true, then we should see roughly the same percentage of people saying the city planner increased levels of joblessness intentionally. However, this is not the case; only 29% of respondents answered that the city planner acted intentionally to increase joblessness levels, which is statistically lower than chance: $\chi^2(1, N = 21) = 3.857, P = .05$ This seems to be clear evidence of the falsity of BAM as it has been recently articulated (Knobe, 2006).⁸ Apprehensions of foreknown, bad side effects are not sufficient to influence judgments of intentionality.

However, in rejecting BAM, we are not thereby endorsing BLAM. Only 8 subjects (38%) thought the city planner deserved blame for having increased joblessness, and of these, only 1 thought he acted intentionally. In other words, subjects who replied that the city planner deserved blame were extremely unlikely to say he acted intentionally.

IV

We had two goals in this paper. First, to show that BAM claimed victory prematurely by failing to use a case that was relevant to the debate. Second, to show that BAM is, in fact, false. In arguing that BAM is false, we are not ruling out that evaluations of side effects influence judgments of intentionality in a way suggesting some conceptual link between the two. Furthermore, as we have just pointed out, our data tentatively suggest that judgments of blameworthiness are not always predictive of intentionality judgments either. It appears that some more complicated story is correct, but just which story awaits further investigation. For now we have been satisfied to critique BAM, and to show that, so long as you weren’t trying, there seem to be situations in which ordinary people think you didn’t act intentionally, even though what you did was bad and you knew about it.

Acknowledgements We would like to thank Jesse Prinz and an anonymous reviewer for *Philosophical Studies* for their illuminating comments on previous drafts. We owe special thanks to Joshua Knobe, who offered essential help throughout the process of researching and writing this paper.

References

- Adams, F. (1986). Intention and intentional action: The simple view, *Mind and Language* 1, 281–301.
- Adams, F. (1997). Cognitive Trying. In G. Holmstron-Hintikka & R. Tuomela (Eds.), *Contemporary Action Theory* (pp. 287–314, vol 1). Dordrecht: Kluwer.
- Davidson, D. (1963). Actions, Reasons, and Causes. In *The Essential Davidson*, Oxford: Oxford University Press, 23–36.

⁸ We shared our results with Knobe, and he agrees (personal communication) that they do, indeed, constitute a refutation of BAM.

- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.
- Harman, G. (1976). Practical reasoning. *Review of Metaphysics*, *29*, 431–463.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190–193.
- Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, *9*, 357–359.
- Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, *130*, 203–231.
- 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.
- Lowe, E. J. (1978). Neither intentional nor unintentional. *Analysis*, *38*, 117–118.
- McCann, H. (1986). Rationality and the range of intention. *Midwest Studies in Philosophy*, *10*, 191–211.
- Mele, A. R., & Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. *Philosophical Studies*, *82*, 265–287.
- Nadelhoffer, T. (2004a). The butler problem revisited. *Analysis*, *64*(3), 277–284.
- Nadelhoffer, T. (2004b). On praise, side effects, and folk ascriptions of intentionality. *Journal of Theoretical and Philosophical Psychology*, *24*, 196–213.
- Nadelhoffer, T. (2004c). Blame, badness, and intentional action: A reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology*, *24*, 259–269.
- Nadelhoffer, T. (2006). Desire, foresight, intentions, and intentional actions: Probing folk intuitions. *Journal of Cognition and Culture*, *24*, 133–157.
- Pitcher, G. (1970). In intending and side effects. *Journal of Philosophy*, *67*, 659–668.