Philosophical Studies (2006), 130, 203-231

The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology

Joshua Knobe
UNC-Chapel Hill

Note: The text presented here is from a slightly revised version that was put together for inclusion in the forthcoming volume *Experimental Philosophy*.

The twentieth century saw the rise of a new discipline that we might call *scientific psychology*. Practitioners of this new discipline develop detailed theories, conduct systematic experiments and publish their results in academic journals.

But long before the rise of scientific psychology, people had ways of making sense of the goings-on in each other's minds. These ordinary ways of understanding the mind did not involve any detailed theories or systematic experiments, but they constituted a kind of psychology all the same. This ordinary, everyday psychology was expressed in sentences like: 'She is feeling angry.' 'He wishes he could go.' 'They think that it is going to rain.' The basic conceptual framework underlying these sorts of everyday psychological ascriptions is usually known as *folk psychology*.

A question now arises about the relationship between folk psychology and scientific psychology. To what extent are they similar, and to what extent different? Over time, researchers working on this question have arrived at a sort of limited consensus. Although considerable disagreement remains about whether or not folk-psychological reasoning actually proceeds using the same kinds of methods one finds in

I am grateful to Gilbert Harman, Bertram Malle, Alfred Mele and Shaun Nichols for comments on an earlier draft.

scientific psychology, almost all researchers now agree that the two kinds of psychology serve more or less the same basic function. Specifically, it is now widely agreed that both kinds of psychology serve primarily to help us predict and explain behavior.

There is something extremely plausible and convincing about the claim that folk psychology plays much the same role in our lives that scientific psychology does. Nonetheless, I think we now have good reason to believe that this claim is not quite right. As I try to show here, certain aspects of folk psychology do not appear to be best understood as tools for predicting and explaining behavior. Instead, these aspects of folk psychology appear to be serving a function that we would never expect to find in a systematic science.

In arguing for this conclusion, I will focus on just one aspect of folk psychology – our folk-psychological concept of *intentional action*. People normally distinguish between behaviors that are performed intentionally (e.g., raising a glass of wine to one's lips) and those that are performed unintentionally (e.g., spilling the wine all over one's shirt). The key question to be addressed here is whether the competencies underlying people's use of this distinction are to be understood primarily in terms of the kinds of aims we normally associate with scientific concepts. I review evidence that indicates that the answer is *no* – i.e., that these competencies have been shaped in a very fundamental way by a quite different sort of function.

By focusing in this way on just one concept, one gains the opportunity for greater depth. That is, one gains the opportunity to examine this one concept in detail and gain real insight into questions about the role it plays in folk psychology. But, of course, to gain this kind of depth, one must sacrifice a certain amount of breadth. It is conceivable (at least in principle) that the concept of intentional action is completely different from every other aspect of folk psychology. Hence, it is conceivable that every other aspect of folk psychology really was shaped almost entirely by its role in 'scientific' tasks (like prediction and explanation) and that the concept of intentional action is the sole exception to this general rule. Although this seems to me to be a somewhat implausible conclusion, I will not be arguing against it explicitly here. The claim is simply that the competencies underlying our folk-psychological concept of intentional action constitute a

counterexample to the view that all of folk psychology should be understood as a device for prediction and explanation.

I

I begin with some straightforward data about people's intuitions concerning specific cases. The key claim here will be that – strange as it may seem – people's intuitions as to whether or not a behavior was performed intentionally can sometimes be influenced by *moral* considerations. That is to say, when people are wondering whether or not a given behavior was performed intentionally, they are sometimes influenced by their beliefs about whether the behavior itself was good or bad. To find evidence for this claim, we can construct pairs of cases that are almost exactly alike except that one involves a harmful behavior and the other a helpful behavior. It can then be shown that these different behaviors elicit different intuitions.

For a simple example, consider the following story:

> 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.

Now ask yourself: Did the chairman of the board *intentionally* harm the environment?

Faced with this question, most people (though certainly not all) say that the answer is yes. And when asked why they think that the chairman intentionally harmed the environment, they tend to mention something about the chairman's psychological state — e.g., that he decided to implement the program even though he specifically knew that he would thereby be harming the environment.

But it seems clear that these facts about the agent's psychological state cannot be all there is to the story. For suppose that we replace the word 'harm' with 'help,' so that the vignette becomes:

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

This one change in the vignette leads to a quite radical change in people's intuitions. Faced with this second version, most people say that the chairman did *not* intentionally help the environment.

To confirm these claims about people's intuitions, I presented the two vignettes to subjects in a controlled experiment (Knobe 2003a). The results were clear and compelling: 82% of subjects who received the story about environmental harm said that the chairman harmed the environment intentionally, whereas only 23% of subjects who received the story about environmental help said that the chairman helped the environment intentionally. This result provides preliminary evidence for the view that people's beliefs about the moral status of a behavior have some influence on their intuitions about whether or not the behavior was performed intentionally.

Of course, it would be a mistake to base such a broad claim on evidence from just one vignette. But the claim becomes plausible when one sees how robust the effect is. The effect continues to emerge when the whole experiment is translated into Hindi and run with Indian subjects (Knobe & Burra 2006); it emerges when subjects are only 4 years old (Leslie et al. 2006); it even emerges when the experiment is run on subjects who suffer deficits in emotional processing due to lesions in the ventromedial prefrontal cortex (Young et al. 2006). Moreover, philosophers have constructed other, very different

cases in which moral considerations appear to influence people's intuitions about whether or not a given behavior is intentional (Harman 1976; Lowe 1978), and when these other kinds of cases have been put to an experimental test, the effect emerges on them as well (Knobe 2003b).

To some degree at least, it seems that these results should come as a surprise to those who think of people's concept of intentional action as a tool for predicting and explaining behavior. After all, it seems that the best way to accomplish these 'scientific' goals would be to ignore all the moral issues and focus entirely on a different sort of question (e.g., on questions about the agent's mental states). How then are we to make sense of the fact that moral considerations sometimes influence people's application of the concept of intentional action?

By now, it should be clear where I am heading. What I want to suggest is that there is another use of the concept of intentional action in light of which the influence of moral considerations really does make sense. The claim is that people's concept of intentional action should not be understood simply as a tool for predicting and explaining behavior. The concept has also been shaped in a very fundamental way by a different kind of use, and it is only by considering this second use that we will be able to reach an adequate understanding of the surprising results we have just described.

II

Before taking up this issue in more detail, let us pause to consider the structure of the cases in which people's intuitions appear to be influenced by moral considerations. Here our aim is simply to amass some useful data about people's intentional action intuitions. We will defer to a later section all questions about *why* people have these intuitions and what these intuitions indicate about the role of intentional action in folk psychology.

In describing the factors that influence people's intuitions, it will often prove helpful to make reference to the various features that philosophers have discussed in their analyses of the concept of intentional action. Here we shall be principally concerned with the features *trying*, *foresight* and *skill*. There has been a great deal of controversy in the philosophical literature about the role that each of these features plays in the concept of

intentional action (for an excellent review, see Mele 1992). In the present context, however, it will not be necessary to discuss these controversies in any real detail. Instead, what we want to show is that, in the cases under dispute, people's intuitions are influenced by the moral status of the behavior.

First, let us consider the debate surrounding the role of *trying* and *foresight*. Some philosophers think that trying is a necessary condition for intentional action (Adams 1986; McCann 1986); others argue that a certain kind of foresight can actually be sufficient even in the absence of trying (Ginet 1990).

The distinction between these two views comes out most clearly in cases of what might be called *side-effects*. An outcome can be considered a 'side-effect' when (1) the agent was not specifically trying to bring it about but (2) the agent chose to do something that she foresaw would involve bringing it about. The question is: Will people think that the agent brought about such an outcome *intentionally*?

An examination of such cases can help us understand the roles played by judgments of trying and foresight in generating people's intentional action intuitions. If people take trying to be a necessary condition, they should think that the agent did not bring about the side-effect intentionally. By contrast, if they take foresight to be sufficient, they should think that the agent did bring about the effect intentionally. But when we study these cases systematically, we end up with a surprising result: people's intuitions appear to be influenced by the *moral* qualities of the side-effect itself. Specifically, people seem to be considerably more willing to say that the agent brought about the side-effect intentionally when they regard that side-effect as bad than when they regard the side-effect as good.

This is the key result of the experiment described above — where a vignette about environmental harm elicited very different intuitions from a quite similar vignette about environmental help. And the same effect arises for other cases that have the same basic structure. So, for example, when we transpose the story from a corporate boardroom to a battlefield — with a lieutenant helping or harming his troops in place of a chairman helping or harming the environment — we still get the same basic effect. People say that the lieutenant acted intentionally if he harmed the troops as a side-effect but that he did not act intentionally if he helped the troops as a side-effect (Knobe 2003a).

Cases of side-effects are not the only ones in which moral considerations play a role. Similar issues arise in cases where the agent lacks *skill*. Consider a case in which an agent is trying to perform a behavior and actually does succeed in performing that behavior. And now suppose that the agent didn't really have the skill to perform that behavior in any reliable fashion, so that ultimately the agent only manages to succeed through sheer luck. Has the agent performed the behavior intentionally? According to some philosophical analyses, the answer is yes (e.g., Brand 1984); according to others, the answer is no (e.g., Mele & Moser 1994). But once again, it appears that neither view correctly predicts people's intuitions in all cases. People's intuitions about these cases seem to depend in part on the moral status of the behavior itself.

Here it may be helpful to consider another series of cases. First, take a case in which the agent's behavior might be regarded as an *achievement*:

> Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bulls-eye. He raises the rifle, gets the bull's-eye in the sights, and presses the trigger.
>
> But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…
>
> Nonetheless, the bullet lands directly on the bull's-eye. Jake wins the contest.

Faced with this case, most people think that it would be wrong to say that Jake hit the bull's-eye intentionally.

But now suppose that we consider a case that is quite similar in certain respects but in which the behavior would normally be regarded as *immoral*:

> Jake desperately wants to have more money. He knows that he will inherit a lot of money when his aunt dies. One day, he sees his aunt walking by the window. He raises his rifle, gets her in the sights, and presses the trigger.

> But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…
>
> Nonetheless, the bullet hits her directly in the heart. She dies instantly.

Changing the moral significance of the behavior in this way leads to a quite substantial change in the pattern of people's intuitions. Faced with this second vignette, people overwhelmingly say that Jake hit his aunt intentionally.

Finally, let us consider a case in which the agent's behavior would normally be seen as *morally good*:

> Klaus is a soldier in the German army during World War II. His regiment has been sent on a mission that he believes to be deeply immoral. He knows that many innocent people will die unless he can somehow stop the mission before it is completed. One day, it occurs to him that the best way to sabotage the mission would be to shoot a bullet into his own regiment's communication device.
>
> He knows that, if he gets caught shooting the device, he may be imprisoned, tortured or even killed. He could try to pretend that he was simply making a mistake — that he just got confused and thought the device belonged to the enemy — but he is almost certain that no one will believe him.
>
> With that thought in mind, he raises his rifle, gets the device in his sights, and presses the trigger. But Klaus isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…
>
> Nonetheless, the bullet lands directly in the communications device. The mission is foiled, and many innocent lives are saved.

Here most people feel that Klaus did hit the communications device intentionally.

In fact, the differences among these vignettes have been demonstrated experimentally — with 23% of subjects saying that Jake intentionally hit the target in the achievement vignette, 91% in the immoral vignette, and 92% in the morally good vignette (Knobe 2003b). Once again, it appears that people's intentional action intuitions are in some way influenced by their beliefs about the moral status of the behavior itself.

Thus far, we have reported results from only two experiments. But these results have been replicated and extended in a considerable body of work by both philosophers and psychologists (Feltz & Cokely 2007; Leslie et al. 2006; Malle 2006; McCann 2005; Nadelhoffer 2004a, 2005; Nichols & Ulatowski forthcoming; Pizarro et al. 2007; Sverdlik 2004; Young et al. 2006). At this point, there can be little doubt that moral considerations have an impact on people's use of the word 'intentionally.' The key remaining questions are about how this effect is to be understood.

III

In particular, a question arises as to whether moral considerations are actually playing a role in the fundamental competencies underlying our use of the concept of intentional action. After all, it is possible that moral considerations could have a decisive impact on our use of words like 'intentionally' even if they have no impact at all in these underlying competencies. Some additional process could be intervening between the underlying competencies and our use of words, and it could be that this additional process is the only place in which moral considerations have a real impact.

Still, it isn't enough just to point out that there might be some other way to explain the findings. What one wants is an alternative model, a specific hypothesis about how an intervening process might be shaping our use of the word 'intentionally' in a way that is more or less unrelated to our underlying competencies. Then we can check to see whether this alternative model gives us a better account of the data than the straightforward hypothesis that moral considerations are playing some role in the competencies themselves.

Of course, it will never be possible to assess all conceivable alternative models. We therefore proceed by considering three models that have actually been proposed.

1. Mele (2001) suggests that the effect might be due, not to people's (largely tacit) concept of intentional action, but rather to certain explicit beliefs they hold about the relation between intentional action and moral blame. Specifically, he suggests that people hold an explicit belief that an agent can only be blameworthy for performing a behavior if that agent performed the behavior intentionally. This explicit belief might be more or less unrelated to the purely tacit mechanisms that normally direct people's application of the concept of intentional action. Indeed, the content of the belief might directly contradict the contents of the non-conscious states that make these mechanisms possible.

Still, the content of people's explicit beliefs could be having a large impact on their responses to specific cases. When they encounter a case like that of the executive harming the environment, their tacit competence might spit out the conclusion: 'This behavior is unintentional.' But then they might think: 'Wait! The agent is clearly to blame for his behavior, and agents can only be blameworthy for performing intentional actions. So the behavior in question just *must* be intentional after all.'

It certainly does seem possible, as Mele suggests, that people hold various explicit beliefs about the relation between intentional action and moral blame. The question is simply whether these explicit beliefs alone can explain all of the ways in which moral considerations appear to be influencing people's application of the concept of intentional action. Suppose, for example, that people somehow ceased to believe that all blameworthy behaviors were intentional. Would moral considerations still continue to have an impact on their application of the concept of intentional action?

To address this question, I tried to create a situation in which people would come to believe that a behavior can be blameworthy even if it is not intentional. Subjects were given a story about an agent who performed a behavior unintentionally but seemed clearly to be deserving of blame. (The story concerned an agent who harms other people while driving drunk.) Subjects were then asked (a) whether or not the agent acted intentionally and (b) whether or not the agent was to blame for his behavior. As expected, almost all subjects answered no to the first question and yes to the second. Immediately after answering this question, subjects were presented with a case in which moral considerations usually have an impact on people's intentional action intuitions.

Consider the position of a subject answering this second question. Presumably, she does not believe that all blameworthy behaviors have to be intentional. (After all, in her answer to the previous question, she said explicitly that the agent acted unintentionally but was blameworthy nonetheless.) She now faces a story about an agent who performed an immoral behavior. The key question is whether the moral status of the agent's behavior will have any impact on her judgment as to whether or not it was performed intentionally.

The answer is that the moral status of the behavior continues to have an impact even in this situation. As in previous studies, subjects were far more likely to classify the behavior as intentional when it was morally bad. Faced with this new result, Mele (2003) has retracted his previous view. He now claims that moral considerations do indeed play a role in people's concept of intentional action.

2. Adams and Steadman (2004a) suggest that the effect might be due entirely to conversational pragmatics. The basic idea is that people are describing blameworthy behaviors as 'intentional' because they want to avoid certain unwanted implicatures. When a person utters the sentence 'He didn't do that intentionally,' there is often a clear implicature that the agent is not to blame for what he has done. Thus, when people are asked whether the chairman harmed the environment intentionally or unintentionally, they may be understandably reluctant to respond that his behavior was entirely unintentional.

The alleged problem here lies in the specific method by which we have been trying to figure out whether people regard a given behavior as intentional. Our method has been to look at people's application of the word 'intentional' and, from that, to make inferences about which behaviors they truly believe to have been performed intentionally. But, as Adams and Steadman rightly point out, people's use of this word is no sure guide to their application of the corresponding concept. Factors like conversational pragmatics may influence people's use of words even if they play no role at all in the fundamental competencies underlying folk psychology.

What we need here, ideally, is some independent method for figuring out whether people regard a given behavior as intentional — a method that makes no use of the word

‘intentionally.’ Then we can check our earlier results against the results obtained using this independent method. If the independent method yields results that differ in some important respect from those obtained when we simply asked people whether a given behavior was performed intentionally, we might suspect that our earlier results were due in part to pragmatic factors and did not truly reveal people’s underlying concept of intentional action. If, however, the alternative method yields the very same results we obtained using the original method, we would have good reason to believe that those earlier results were telling us something important about which behaviors people truly regard as intentional.

As it happens, there is such an independent method. We can determine whether or not people regard a given behavior as intentional by looking at their use of the phrase ‘in order to.’ It seems that people are generally unwilling to say that an agent performed a behavior ‘in order to’ attain a particular goal unless they believe that the agent performed that behavior intentionally. Thus, if a speaker utters a sentence of the form ‘She *A*-ed in order to *B*,’ we would normally assume that the speaker takes the agent to have *A*-ed intentionally.

Using this alternative method, we can retest our original hypothesis. Do people genuinely regard the harming of the environment as an intentional action, or are they only labeling it ‘intentional’ because they want to avoid certain pragmatic implicatures? One way to find out would be to ask whether people are willing to say that the chairman harmed the environment ‘in order to’ attain a particular goal. In actual fact, it appears that they regard some sentences of this form as perfectly acceptable. Faced with the harm vignette, people generally think it sounds right to say:

‘The chairman harmed the environment in order to increase profits.’

But, surprisingly enough, people who have been given the help vignette do not generally think it sounds right to say:

‘The chairman helped the environment in order to increase profits.’

Presumably, this asymmetry in people’s use of the phrase ‘in order to’ reflects an asymmetry in people’s views about which behaviors were performed intentionally (Knobe 2004). Since people regard the harming of the environment as intentional and the

helping of the environment as unintentional, they are willing to use the phrase ‘in order to’ for harming but not for helping.

Adams and Steadman (2004b) are not convinced by this response. They argue that the effect for ‘in order to’ can be understood in terms of the very same pragmatic processes they had originally posited to explain the effect for ‘intentionally.’ The idea is that people see immediately that no agent can perform a behavior ‘in order to’ attain a goal unless that agent performs the behavior intentionally. Any factor that has an impact on the pragmatics of ‘intentionally’ should therefore have an impact on the pragmatics of ‘in order to’ as well.

Although Adams and Steadman may ultimately turn out to be right on this score, their pragmatic explanation for the use of ‘in order to’ definitely lacks the intuitive plausibility of the explanation they originally offered for the use of ‘intentionally.’ It is common practice to deny that an agent deserves blame by saying ‘He didn’t do that intentionally,’ but we do not normally deny that an agent is blameworthy by using a sentence like ‘It doesn’t sound right to say that he did that “in order to” attain a goal.’ In fact, if someone used such a sentence in an ordinary conversation, we would probably have no idea what she was trying to say. There seems not to be any direct connection between being blameless and not performing an action in order to attain a goal. The only way to recover the alleged implicature here would be to first (a) infer that the use of ‘in order to’ was sounding wrong because the behavior itself was unintentional, then (b) determine that classifying a behavior as unintentional indicates that the behavior is not deserving of blame, and finally (c) conclude that the sentence therefore implicates that the agent is not blameworthy. Such a complex chain of reasoning could hardly take place in the few seconds it normally takes people to answer these questions.

3. Nadelhoffer (2004b) and Malle and Nelson (Malle & Nelson 2003; cf. Malle 2004) suggest that the data are best explained in terms of the distorting effects of people’s feelings of *blame*. The key idea here is that moral considerations play no role at all in the fundamental competence underlying people’s concept of intentional action. However, when people classify an agent’s behavior as immoral, they may quickly come to feel that the agent is deserving of blame. This feeling then distorts their reasoning, leaving them

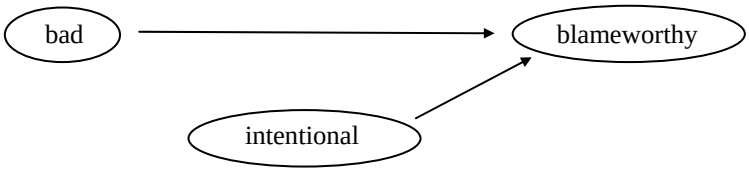
with a strong motivation to declare the agent's behavior intentional and thereby justify the blame they have already assigned.

Before evaluating this hypothesis in more detail, we need to make a few preliminary comments about the notion of moral blame itself. Then we can compare a number of competing models of the relationship between judgments of blame and the concept of intentional action. The aim will be to see which of these models best explains people's intuitions about specific cases.

To begin with, we need to make a clear distinction between the judgment that a behavior is *bad* and the judgment that an agent is *blameworthy*. Consider the agent who hurts his wife's feelings. Here we might say that the agent's behavior itself is bad. That is to say, when we ignore every other aspect of the situation, we might classify the hurting of the wife's feelings as a bad thing. Still, we will be unlikely to blame the agent if he has a good excuse (ignorance, mental illness, provocation, etc.) or if his behavior is in some way justified (e.g., because hurting his wife's feelings leads to some good consequence in the long run).

These two kinds of judgments seem to result from two distinct stages in the process of moral assessment. First we make a judgment as to whether or not the behavior itself is bad and then — depending on the outcome of this first stage — we may end up making a judgment as to whether or not the agent deserves blame. Where in this whole process does the concept of intentional action appear?

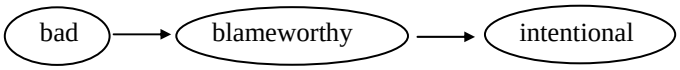
The commonsense view works something like this:



On this model, people determine whether the behavior itself is bad without making any use of the concept of intentional action. However, they do use the concept of intentional action when they are trying to determine whether or not the agent deserves blame.

One problem with this commonsense view is that it offers no explanation for the fact that people's moral judgments sometimes influence their intuitions as to whether or

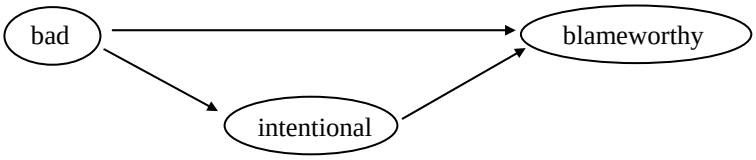
not a behavior was performed intentionally. Nadelhoffer, Malle and Nelson therefore propose that the process sometimes works more like this:



On this model, people do not use the concept of intentional action to determine whether or not the agent is blameworthy. Instead, they assign blame *before* they have even applied the concept. Then they apply the concept in such a way as to justify the blame they have already assigned.

If the process really does work like this, it would be reasonable to infer that people were making some kind of error. This model does not posit a role for moral considerations in the fundamental competence underlying people's concept of intentional action. Rather it seems to be describing a kind of bias that can infect people's thought processes and lead them astray.

There is, however, another plausible way to make sense of the data reported thus far. Perhaps the process actually works like this:



This third model can make sense of the fact that people's moral judgments sometimes influence their intuitions as to whether or not a behavior was performed intentionally, but it also retains the commonsense view that people use the concept of intentional action when they are trying to determine whether or not the agent deserves blame. The basic idea is that people's judgment that the behavior itself is bad can influence their intuitions as to whether the behavior was performed intentionally and that these intuitions can, in turn, play an important role in the process by which people determine whether or not to assign blame.

In the cases we have been discussing thus far, these competing models make identical predictions. Take the case of the corporate executive who harms the environment. Here we find that people both (a) classify the agent's behavior as bad and

(b) blame the agent for that behavior. Since people judge the case to be both bad and blameworthy, there is no obvious way to figure out which of these two judgments is influencing their intuitions.

To decide between the competing models, we therefore need to find a case in which an agent brings about a bad side-effect but is *not* considered blameworthy. In such a case, the different models will yield different predictions. If the badness of the side-effect only impacts people's intuitions by first leading to feelings of blame, people should be inclined to regard the side-effect as unintentional. But if people's intuitions can be directly influenced by judgments of badness — without any mediation of feelings of blame — they should be inclined to regard the side-effect as intentional.

For a simple test case, let us modify our vignette about the corporate executive trying to decide whether or not to implement a new program. This time, we will not suppose that the program leads to environmental harm or any other morally significant consequence. Instead, we can suppose that the program has only two important effects: it increases sales in Massachusetts but decreases sales in New Jersey. The executive knows that the gain in Massachusetts will be far larger than the loss in New Jersey, and she therefore decides to implement the program.

Now consider the status of the behavior *decreasing sales in New Jersey*. Here it seems that the agent has done something bad without being in any way blameworthy. When we say that the agent's behavior is bad, we simply mean that decreasing sales in New Jersey is, taken in itself, a bad thing. Of course, it isn't *morally* bad to decrease sales, and it might even be helpful on the whole, given its consequences. Still, there is a straightforward sense in which one might say: 'It's *too bad* that she had to decrease sales in New Jersey.' At the same time, though, it is clear that the agent is in no way deserving of blame for her behavior. If anything, she deserves praise for finding a policy that increases sales on the whole.

And yet, people generally say that the executive intentionally decreased sales in New Jersey (Knobe & Mendlow 2004). This result spells trouble for any theory that tries to account for the role of moral considerations in terms of blame alone. What we have here is a case in which the agent is not considered blameworthy but in which people's beliefs about good and bad are nonetheless influencing their intentional action intuitions.

This kind of result cannot plausibly be explained in terms of people's efforts to justify a prior judgment of blame. (After all, there is no blame here to justify!) The most plausible hypothesis seems to be that people's judgments of good and bad are actually playing a role in the fundamental competencies underlying their concept of intentional action.[1]

Thus far, we have been considering the evidence for and against specific alternative models. Ultimately, though, it may not be enough merely to consider the various alternative models that are already available in the literature. No matter how many alternative models one eliminates, it will always be possible for future researchers to devise new ones. Indeed, even in the absence of any specific alternative model, one may be tempted to suppose that *some* alternative model can adequately explain the data. What we need to address, then, is the widespread sense — never explicitly defended but deeply felt nonetheless — that an alternative model is needed. That is to say, we need to address the widespread sense that moral considerations just *couldn't* be playing any role in the fundamental competencies underlying folk psychology.

This sense is never fully articulated by any of the authors cited above. Instead of arguing explicitly against the view that moral considerations play some fundamental role in folk psychology, these authors simply propose alternative models and then try to show that their models provide plausible explanations of the data. The presumption seems to be that, if any alternative model can provide a plausible explanation, that model is to be preferred over the hypothesis that moral considerations really are playing a role in folk psychology. But what is the source of this presumption?

The answer lies, I think, in a particular view about the nature of folk psychology. This view says that the basic purpose of folk psychology is to enable people to predict each other's behavior or to offer them some other form of quasi-scientific, purely

[1] [*Note added 2007*] The hypothesis offered in this early paper is that the only type of moral judgment that influences people's intentional action intuitions is the judgment that a behavior is *bad*. In the years since I first put forward this hypothesis, it has been put to the test in a number of carefully designed empirical studies (Cushman 2007; Phelan & Sarkissian forthcoming; Tannenbaum, et al. 2007; Sinnott-Armstrong, et al. 2007; Wright & Bengson forthcoming). Sadly, those studies have conclusively demonstrated that my hypothesis was false. The collapse of this original hypothesis has led to a profusion of new models which aim to accommodate all of the recent data while also evading the problems that beset the models I discuss here (e.g., Alicke forthcoming; Knobe forthcoming; Machery forthcoming; Malle 2006; Nadelhoffer 2006; Nichols and Ulatowski forthcoming).

naturalistic understanding. When folk psychology is understood in this way, it seems that it would be *pointless* for moral considerations to play any real role. Thus, if moral considerations appear to be influencing people's use of words like 'intentional,' one is naturally led to search for some alternative to the view that these considerations are actually having an impact in the fundamental competencies underlying folk psychology. The goal then becomes to find some way in which people's fundamental competencies can be overridden, corrupted or otherwise shielded from view.

But, of course, there is another possible approach. Instead of starting out with certain preconceptions about the nature of folk psychology and then trying to square the data with those preconceptions, we can start out with the data and try to figure out what the data might be telling us about the nature of folk psychology. The use of moral considerations may not facilitate the process of predicting behavior, but perhaps we can find some other activity in which the use of moral considerations would prove genuinely helpful.

IV

In particular, let us focus on the process by which people assign praise and blame. It seems clear that the concept of intentional action plays an important role in this process. Specifically, it seems that people are generally inclined to give an agent more praise and blame for behaviors that they regard as intentional than for those they regard as unintentional.

Now suppose that we think of the concept of intentional action in terms of this second use. Suppose, in other words, that we think of it as a tool used for determining how much praise or blame an agent deserves for her behaviors (Bratman 1984; 1987). Then we can check to see whether the criteria according to which people apply the concept seem to make more sense under this construal than they did when we tried to understand every aspect of the concept solely in terms of its 'scientific' use.

First of all, we should note that the three features we encountered in our discussion of intentional action — trying, foresight and skill — play a crucial role in the process by which people normally assign praise and blame. Thus, when people are wondering how much praise or blame an agent deserves, their conclusion will sometimes

depend on whether or not the person was *trying* to perform a given behavior, whether she chose to do something that she *foresaw* would involve performing that behavior, whether she had the *skill* to perform that behavior reliably.

A question now arises as to how people employ information about these various features in making an overall judgment about how much praise or blame the agent deserves. One sees immediately that this process must be extremely complex. It is not as though, e.g., the presence of foresight always increases praise or blame by a constant amount. Rather, different features will be relevant to different behaviors — with a single feature sometimes making a big difference in how much praise or blame an agent gets for one type of behavior yet having almost no impact on the amount of praise or blame that an agent gets for some other type of behavior.

This phenomenon has important implications for the study of praise and blame. It indicates that there is no single way of combining information about psychological features that can be used to determine praise and blame for all possible behaviors. So, for example, suppose we had a concept *shmintentional* that could be given some simple definition like:

> A behavior is shmintentional if and only if the agent had skill and either trying or foresight.

We could not make praise and blame judgments by simply checking to see whether a given behavior was shmintentional. The problem is that different features are relevant to different behaviors and that shmintentionality is therefore more relevant to praise and blame judgments for some behaviors than for others.

For a simple example, we can return to the environmental cases that we presented above. Suppose that an agent decides to perform a given behavior because he wants to increase profits. The agent knows that his behavior will have some impact on the environment. But he does not care at all about the impact he is having on the environment — he is only performing the behavior as a way of increasing profits. Will people feel that this agent deserves any praise or blame for what he has done? Clearly, people's views will depend on the particular type of impact that the agent is having on the environment. If the agent is *harming* the environment, they may feel that he deserves a

considerable amount of blame. But if he is *helping* the environment, they will probably feel that he deserves almost no praise.

What we see here is a remarkable convergence between the conditions under which people assign praise and blame and the conditions under which they regard a behavior as intentional. We noted above that there is a puzzling asymmetry in people's intuitions about intentional action in side-effects cases. People seem to be far more inclined to say that an agent brought about a side-effect intentionally when they regard that side-effect as bad than when they regard it as good. And now we see an analogous asymmetry in people's judgments about praise and blame — namely, that people are far more inclined to give the agent praise or blame for a side-effect when they regard that side-effect as bad than when they regard it as good.

Interestingly, a similar effect emerges for the various cases we described in which the agent lacks the skill to reliably perform the behavior. First, consider the 'achievement' case, where the agent is shooting at a bull's-eye target. There, the amount of praise we give the agent appears to depend on skill, with the agent getting very little praise if his success is due almost entirely to luck. (Our concern here is not with *moral* praise — but we are dealing with a form of praise all the same.) But suppose we consider cases in which the hitting of the target is either immoral or morally good. Then people will tend to give the agent a large amount of praise and blame even when the agent has almost no skill and only manages to hit the target through luck.

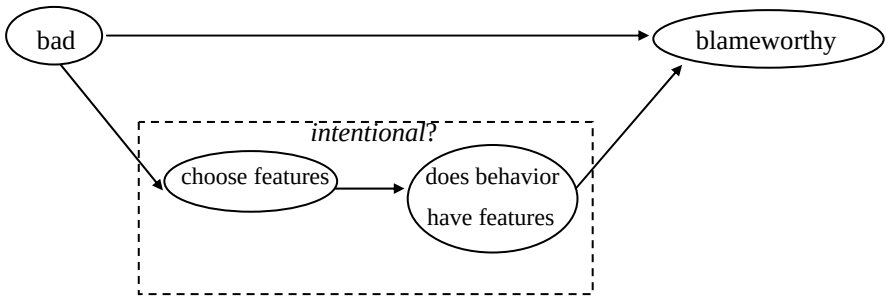
Once again, we find a surprising convergence between people's judgments of praise and blame and their intentional action intuitions. We showed above that people are considerably more likely to say that the hitting of the target is intentional when they regard it either as immoral or as morally good than when they regard it as an achievement. Now we find that this same pattern emerges in people's judgments of praise and blame: people generally give the agent considerably more praise and blame for 'lucky successes' when they regard those successes as immoral or morally good than when they regard them as achievements.

Seen in this light, the pattern of people's intentional action intuitions no longer seems so incoherent or pointless. We have been assuming that people sometimes use the concept of intentional action as a tool for determining how much praise or blame an agent

deserves — with people generally giving the agent more praise and blame for behaviors that they regard as intentional than for behaviors that they regard as unintentional. But we also found that there is no fixed list of features that people always regard as necessary and sufficient for the agent to receive praise or blame for a given behavior. Rather, a given feature may be highly relevant to the praise or blame an agent receives for one behavior while remaining almost entirely irrelevant to the praise or blame the agent receives for another, somewhat different behavior. Thus, if the concept of intentional action is to be helpful in the process of assessing praise and blame, people cannot go about determining whether or not a behavior is intentional by simply checking to see whether it has all the features on some fixed list. People would have to look for different features when confronted with different behaviors. And that seems to be exactly what people do. People's intentional action intuitions seem to exhibit a certain flexibility, such that they look for different features when confronted with different behaviors, and they tend to consider in each case the specific features that would be relevant to determining whether the agent is deserving of praise or blame.

We are now in a position to offer a new hypothesis about the role of moral considerations in people's concept of intentional action. The key claim will be that people's intentional action intuitions tend to track the psychological features that are most relevant to praise and blame judgments. But — and this is where moral considerations come in — different psychological features will be relevant depending on whether the behavior itself is good or bad. That is to say, we use different psychological features when we are (a) trying to determine whether or not an agent deserves blame for her bad behaviors from the ones we use when we are (b) trying to determine whether or not an agent deserves praise for her good behaviors.

We can now offer a somewhat more detailed model than the one presented above.



Here the overall process of determining whether or not the behavior was performed intentionally is broken down into two sub-processes. The first sub-process takes in information about whether the behavior itself is good or bad and uses this information to determine which features are relevant. The second sub-process then checks to see whether the behavior in question actually has these features and thereby generates an intentional action intuition.

Thus, suppose that the person is confronted with the behavior *harming the environment*. The first sub-process might determine that, since the behavior itself is bad, it should be considered intentional if the agent showed either trying or foresight. Then the second sub-process might determine that the agent actually did show foresight and that his behavior is therefore rightly considered intentional.

The chief contribution of this new model is the distinctive status it accords to moral considerations. Gone is the idea that moral considerations are 'distorting' or 'biasing' a process whose real purpose lies elsewhere. Instead, the claim is that moral considerations are playing a helpful role in people's underlying competence itself. They make it possible for people to generate intentional action intuitions that prove helpful in the subsequent process of assessing praise and blame.

V

Folk psychology is widely regarded as a tool for the prediction and explanation of behavior. Since people's concept of intentional action appears to be an integral part of folk psychology, one might be tempted to draw the conclusion that the concept of intentional action should be understood primarily in terms of this 'scientific' use. We have been sketching a theory according to which this conclusion is false. The theory emphasizes instead that the concept of intentional action is used in the process by which people assign praise and blame.

In saying this, we in no way deny that the concept of intentional action is often *used* in the tasks of prediction and explanation. Nor do we deny that it is *adequate* for these tasks — that it can do a decent job of fulfilling various scientific purposes. What we are denying is that the concept is in any sense *specialized* for these tasks.

Instead, it appears that people's concept of intentional action should be understood as something like a multi-purpose tool. If we want to understand why the concept works the way it does, it is not enough to examine its use in the tasks of prediction and explanation. Many important facts about the concept can only be correctly understood when we see that it also plays an important role in the process by which people determine how much praise or blame an agent deserves for his or her behavior.

A question now arises as to how this finding about people's concept of intentional action should affect our views about the nature of folk psychology as a whole. One possibility would be that people's concept of intentional action is simply an exception. That is, it might turn out that all the rest of folk psychology truly is best understood as a collection of tools for predicting and explaining behavior and that the concept of intentional action just happens to be one case in which this otherwise accurate theory breaks down. A second possibility, however, would be that many aspects of folk psychology are susceptible to an analysis like the one we have provided here for the concept of intentional action. In other words, it might turn out that many other aspects of folk psychology are shaped in some important respect by a concern for issues of praise and blame. Such an analysis might be correct for certain trait concepts; it might be correct for our practice of giving reason explanations; it might even be correct for ordinary causal attributions. But these questions lie outside the scope of the present paper. With any luck, they will be addressed in future research.

References

Adams, F. (1986). Intention and Intentional Action: The Simple View. *Mind and Language*, 1, 281-301.

Adams, F. and Steadman, A. (2004a) Intentional Action in Ordinary Language: Core Concept or Pragmatic Understanding? *Analysis*, 64, 173-181

Adams, F. and Steadman, A. (2004b) Intentional Actions and Moral Considerations: Still Pragmatic. *Analysis*, 64, 268-276.

Alicke, M. (forthcoming). Blaming Badly. *Journal of Cognition and Culture*.

Brand, M. (1984). *Intending and Acting*. Cambridge, MA: MIT Press.

Bratman, M. (1984). Two Faces of Intention. *Philosophical Review*, 93, 375-405.

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

Cushman, F. (2007). The effect of moral judgment on causal and intentional attribution: What we say, or how we think? Unpublished manuscript. Harvard University.

Feltz, A. & Cokely, E. (2007). An Anomaly in Intentional Action Ascription: More Evidence of Folk Diversity. *Proceedings of the Cognitive Science Society*.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.

Harman, G. (1976). Practical Reasoning. *Review of Metaphysics*, *29*, 431-463.

Knobe, J. (forthcoming). Reason Explanation in Folk Psychology. *Midwest Studies in Philosophy*.

Knobe, J. (2004). Intention, Intentional Action and Moral Considerations. *Analysis*, *64*, 181-187.

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. & Burra, A. (2006). 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 Considerations 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.

Lowe, E. J. (1978). Neither Intentional nor Unintentional. *Analysis, 38,* 117-118.

Machery, E. (forthcoming). Understanding the Folk Concept of Intentional Action: Philosophical and Experimental Issues. *Mind and Language.*

Malle, B. (2006). Intentionality, Morality, and their Relationship in Human Judgment. *Journal of Cognition and Culture, 6,* 87-113.

Malle, B. F. (2004). *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction.* Cambridge, MA: MIT Press.

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.

McCann, H. (1986). Rationality and the Range of Intention. *Midwest Studies in Philosophy, 10,* 191-211.

McCann, H. (2005). Intentional Action and Intending: Recent Empirical Studies. *Philosophical Psychology,* 18, 737-748.

Mele, A. (1992). Recent Work on Intentional Action. *American Philosophical Quarterly* 29: 199-217.

Mele, A. (2001). Acting Intentionally: Probing Folk Notions. In B. F., Malle, L. J. Moses, & D. Baldwin (Eds), *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: M. I. T. Press.

Mele, A. (2003). Intentional Action: Controversies, Data, and Core Hypotheses. *Philosophical Psychology,* 16, 325-340.

Mele, A. R. & Moser, P. K. (1994). Intentional Action. *Nous, 28,* 39-68.

Nadelhoffer, T. (2004a). The Butler Problem Revisited. *Analysis. 64,* 277-284.

Nadelhoffer, T. (2004b). Praise, Side Effects, and Intentional Action. *The Journal of Theoretical and Philosophical Psychology 24,* 196-213.

Nadelhoffer, T. (2005). Skill, Luck, Control, 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. & Ulatowski, J. (forthcoming). Intuitions and Individual Differences: The Knobe Effect Revisited. *Mind and Language.*

Phelan, M. & Sarkissian, H. (forthcoming). The Folk Strike Back; Or, Why You Didn't Do It Intentionally, Though It Was Bad and You Knew It. *Philosophical Studies*.

Pizarro, D., Knobe, J. & Bloom, P. (2007). College Students Implicitly Judge Interracial Sex and Gay Sex to be Morally Wrong. Unpublished manuscript. Cornell University.

Sinnott-Armstrong, W., Mallon, R., McCoy, T. & Hull, J. (2007). Intention, Temporal Order, and Moral Judgments. Unpublished manuscript. Dartmouth College.

Sverdlik, S. (2004). Intentionality and Moral Judgments in Commonsense Thought about Action. *Journal of Theoretical and Philosophical Psychology 24*, 224-236.

Tannenbaum, D., Ditto, P.H., & Pizarro, D.A. (2007). Different Moral Values Produce Different Judgments of Intentional Action. Unpublished manuscript. University of California-Irvine.

Wright, J. & Bengson, J. (2007). Asymmetries in Folk Judgments of Responsibility and Intentional Action. Unpublished manuscript. University of Wyoming.

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, 291-304.