Folk Psychology and Folk Morality:
Response to Critics[1]

Joshua Knobe
Princeton University

It is often implied, and sometimes explicitly asserted, that folk psychology is best understood as a kind of predictive device. The key contention of this widely held view is that people apply folk-psychological concepts because the application of these concepts enables them to predict future behavior. If we know what an agent believes, desires, intends, etc., we can make a pretty good guess about what he or she will do next.

It seems to me that this picture is not quite right. In a series of recent papers, my colleagues and I have presented data that suggests that *moral* considerations actually play an important role in folk psychology (Knobe 2003a; 2003b; 2004; Knobe & Burra forthcoming; Knobe & Mendlow forthcoming). These findings do not sit well with the view according to which folk psychology is best understood as a predictive device. It appears that folk psychology might be better understood as a kind of multi-purpose tool. It is used not only in making predictive judgments but also in making moral judgments, and both of these uses appear to have shaped the fundamental competencies that underlie it.

One of the most important forms of evidence in this debate comes from studies of the distinction people draw between *intentional* and *unintentional* behavior. These studies indicate that people's intuitions as to whether or not a behavior was performed intentionally can be influenced by their beliefs about the moral status of the behavior

[1] I am grateful to Stephen Butterfill for a question that inspired the experiment reported in section 3. Many of the ideas in section 2 first emerged in a discussion with Shaun Nichols and Josh Greene.

itself (e.g., Knobe 2003a). The key results from these studies have been carefully presented by my critics, and I will not be reviewing them again here.

At times, the contributions of my critics can be fairly technical — dealing as they do with detailed questions about the concept of intentional action rather than with broader questions about the relationship between folk psychology and folk morality. Nonetheless, I think that the criticisms raise a number of important questions, and I hope to show that the technical insights of my critics actually bear on issues of quite broad importance that go beyond the concept of intentional action specifically.

Along the way, I point to a number of unsolved problems. I present some preliminary ideas about how to address a few of these problems, but others elude me entirely. With any luck, future research will uncover methods that enable us to get a handle on them.

1. Most researchers have assumed that the surprising asymmetries observed in recent studies are to be explained in terms of people's moral beliefs. **Jason Turner** suggests an alternative model. Perhaps people's intuitions are not affected by *their own* moral values. Rather, people's intuitions might be affected by their beliefs about the *agent's* values.

Turner is certainly right that this hypothesis is not ruled out by existing studies. In all of those studies, the 'morally bad' behaviors were also behaviors that the agent believed she had some reason not to perform. So, for example, in almost all cases, the agent would be presumed to know that other people would regard her behavior as bad, and subjects might reasonably have supposed that she felt she had some reason to avoid being regarded in this way.

Turner rightly suggests that the best way to test his hypothesis would be to construct a case in which (a) people do not regard the agent's behavior as bad but (b) the agent does feel that she has a reason not to perform the behavior. It will then be possible to determine which sort of belief is truly driving the observed effect — beliefs about the true goodness or badness of the behavior or beliefs about what the agent thinks she has reason to do.

Consider now the following vignette:

> A terrorist discovers that someone has planted a bomb in a nightclub. There are lots of Americans in the nightclub who will be injured or killed if the bomb goes off. The terrorist says to himself, "Whoever planted that bomb in the nightclub did a good thing. Americans are evil! The world will be a better place when more of them are injured or dead."
>
> Later, the terrorist discovers that his only son, whom he loves dearly, is in the nightclub as well. If the bomb goes off, his son will certainly be injured or killed. The terrorist then says to himself, "The only way I can save my son is to defuse the bomb. But if I defuse the bomb, I'll be saving those evil Americans as well... What should I do?"
>
> After carefully considering the matter, he thinks to himself, "I know it is wrong to save Americans, but I can't rescue my son without saving those Americans as well. I guess I'll just have to defuse the bomb."
>
> He defuses the bomb, and all of the Americans are saved.

Now consider the status of the side-effect *saving the Americans*. Presumably, most readers of the vignette will not regard this side-effect as bad. (In fact, they will almost certainly regard it as good.) Still, it is clear that the agent himself thinks he has strong reason to avoid bringing the side-effect about.

Turner's hypothesis therefore generates a prediction that differs from that of most previous researchers. Whereas previous researchers would have predicted that people would take the side-effect to be unintentional (because they do not regard it as bad), Turner should predict that people will take the side-effect to be intentional (since they know that the agent thinks he has strong reason not to bring it about).

As it happens, when Sean Kelly and I presented this vignette to subjects, most took the side-effect to be unintentional (Knobe & Kelly 2004). This result seems extremely problematic for any view according to which people's judgments are based entirely on their beliefs about the agent's values rather than their beliefs about whether the behavior really is good or bad.

Still, there may be some truth to Turner's hypothesis. Perhaps if the agent's values are made sufficiently salient, people's judgments will be determined by those values rather than by their own moral beliefs. This prediction awaits further empirical study.

2. When we see that moral considerations have some impact on people's intuitions as to whether or not a behavior was performed intentionally, we needn't thereby conclude that moral considerations actually play any role in people's *concept* of intentional action. After all, it is always possible that moral considerations are merely 'distorting' people's intuitions or otherwise getting in the way of the criteria specified by the underlying concept. A number of researchers have tried to explain the recent experimental results by putting forth hypotheses of this basic form (Adams & Steadman 2004; forthcoming; Malle forthcoming; Malle & Nelson 2003; Nadelhoffer forthcoming). All of these

hypotheses suggest that people's intuitions are in some way affected by their feelings of blame.

Arguing against a key presupposition of these various hypotheses, Gabriel Mendlow and I (forthcoming) claimed that feelings of *blame* have little or no impact on people's intuitions. Instead, we suggested that people's intuitions were affected by their beliefs as to whether the behavior itself was *bad*. (Thus, people's intuitions would be affected any time they classified the behavior itself as bad — even in those cases where they felt that the agent was in no way deserving of blame.) Such an effect would not be predicted by any of the models according to which the impact of moral considerations was merely a 'distortion.'

Thomas Nadelhoffer now points out that Mendlow and I never really provided any evidence for the claim that blame had no effect. All we showed was that badness *did* have an effect. But it is surely possible that badness and blame *both* have some impact here.

Nadelhoffer is clearly right on this score. Our experiment definitely does not show that blame has no impact on people's intuitions. (In fact, it's hard to see how any experiment could provide evidence for the claim that blame never has any impact on people's intentional action intuitions.) So what we really should have said was that our experiment showed that the badness of the behavior itself can have an impact.

To show that blame also has an impact, one would have to construct a pair of vignettes that were almost exactly alike but that differed in the degree to which they elicited blame. For example, the two vignettes could be constructed in such a way that they differed only in the background information given about the agent. One agent would

be described as lovable, the other as despicable, but the two vignettes would be exactly the same in all other respects — even in the beliefs and desires that the agent is described as holding with regard to the behavior in question. If this manipulation appeared to be affecting people's intentional action intuitions, we would have good evidence for the view that blame really did have some impact.

On some level, though, it seems that this whole question is more or less unrelated to the key philosophical issue. Everyone already agrees that people's feelings can sometimes distort their folk-psychological judgments; the key question was whether people's moral beliefs also play a role in the basic competencies underlying folk psychology.[2] The existing evidence seems to suggest that people's beliefs about the goodness or badness of a behavior do play such a role. If we later discover that people's intuitions are sometimes distorted by their feelings of blame, this conclusion will remain untouched.

3. **Julie Yoo** approaches this issue from a somewhat different perspective. She points out that my earlier experiments involved behaviors that were *morally* bad (Knobe 2003a, 2003b). Clearly, my recent work with Mendlow is a departure from this tradition. Our experiment does involve a behavior that can in some sense be considered 'bad,' but the badness of this behavior has nothing to do with morality. (It is a purely extra-moral kind of badness that involves making decisions that decrease corporate sales.) Yoo thinks that,

[2] Nadelhoffer has published a number of important papers on this issue, all of them arguing for the view that moral considerations really do play a role in the very concept of intentional action (Nadelhoffer forthcoming a; forthcoming b; forthcoming c). However, I am not sure if I understand his present position. He suggests that moral considerations must figure in a correct analysis of the concept of intentional action, but he also describes the impact of moral considerations as a 'bias' that can be explained in terms of Mark Alicke's theory of blame validation. These two views do not seem to me to be consistent. If the impact of moral considerations is merely a 'bias,' why should moral considerations figure in the correct analysis of the concept?

by departing in this way from the established tradition, we fail to address the questions that arose out of earlier experiments. After all, it is possible that extra-moral evaluations simply don't affect people's intuitions in the same way that moral evaluations do. Thus, it might be a mistake to assume that the findings obtained in experiments with extra-moral evaluation will hold equally well in cases of moral evaluation.

Yoo is drawing our attention to an important question here. Is there something special about moral evaluations that allows them to have a unique effect on people's use of the concept of intentional action? Or do all evaluations affect people's use of the concept in the same way? For example, suppose we describe two behaviors that differ only in their *aesthetic* status — with one behavior making things aesthetically better and the other making things aesthetically worse. Would this aesthetic difference affect people's intuitions in the same way that moral differences usually do?

To address this question, I ran an additional experiment. Subjects were 54 people spending time in a Manhattan public park. Each subject was randomly assigned either to the 'aesthetic harm' condition or to the 'aesthetic help' condition. Subjects in the aesthetic harm condition were asked to read the following vignette:

> The Vice-President of a movie studio was talking with the CEO. The Vice-President said: "We are thinking of implementing a new policy. If we implement the policy, it will definitely increase profits for our corporation, but it will also make our movies worse from an artistic standpoint."
>
> The CEO said: "Look, I know that we'll be making the movies worse from an artistic standpoint, but I don't care one bit about that. All I care about is making as much profit as I can. Let's implement the new policy!"
>
> They implemented the policy. As expected, the policy made the movies worse from an artistic standpoint.

Subjects in the aesthetic help condition read a vignette that was almost exactly the same, except that the word 'worse' was changed to 'better':

> The Vice-President of a movie studio was talking with the CEO. The Vice-President said: "We are thinking of implementing a new policy. If we implement the policy, it will definitely increase profits for our corporation, and it will also make our movies better from an artistic standpoint."
>
> The CEO said: "Look, I know that we'll be making the movies better from an artistic standpoint, but I don't care one bit about that. All I care about is making as much profit as I can. Let's implement the new policy!"
>
> They implemented the policy. As expected, the policy made the movies better from an artistic standpoint.

After reading the assigned vignette, all subjects received two questions. The first question was: 'Did the CEO *intentionally* make the movies worse [better] from an artistic standpoint?' The second question was: 'How much blame or praise does the CEO deserve for what he did?' This second question was answered on a scale from -3 ('a lot of blame') to +3 ('a lot of praise'), with the 0 point marked 'no praise or blame.'

The mean rating for blame or praise in the aesthetic harm condition was -1.7; the mean in the aesthetic help condition was .3. This difference was statistically significant, $t(54) = 5.8$, $p < .001$.

The key question, however, was whether subjects would say that the agent acted intentionally. Here again, there was a statistically significant difference. Approximately half (54%) of subjects in the aesthetic harm condition said that the agent acted intentionally, whereas relatively few (18%) of subjects in the aesthetic help condition said that the agent acted intentionally, $\chi^2(1, N=54) = 7.7$, $p < .01$. In short, it appears that aesthetic evaluations had the same kind of effect that moral evaluations usually do,

although the size of the effect in our aesthetic case was substantially smaller than the effect normally observed in moral cases.

Our findings therefore lend some initial support to the hypothesis that extra-moral evaluations affect people's intuitions in the same way that moral evaluations do. But Yoo's objection still stands. That is to say, it still remains to be seen whether *all* evaluations affect people's intuitions in the same way or whether different kinds of evaluations can have different effects. For example, would a purely grammatical evaluation — a judgment about whether a sentence was grammatically correct or incorrect — affect people's intuitions in the same way that moral judgments do? The only way to resolve such questions is through further experimentation.

4. In my earlier studies, subjects were always forced to choose between two options. The question was whether or not a behavior was performed intentionally, and the only permissible answers were 'yes' and 'no.' **Roblin Meeks** suggests that these two options may not have been sufficient. Perhaps certain behaviors are classified neither as intentional nor as unintentional. Presented with such a behavior, subjects might be reluctant to give either of the two usual answers.

Meeks is right to emphasize that some behaviors don't fit comfortably into either category. For a clear example, consider the following story, which I will call the *Lottery Vignette*:

> Jake wanted to win the lottery. So he decided to buy a ticket.
>
> He knew he only had a 1 in 500,000 chance of winning, but he thought that it might be a good idea to give it a try anyway.

Jake was extremely lucky. He ended up winning the lottery.

Here we would presumably be reluctant to say either that 'Jake intentionally won the lottery' or that 'Jake unintentionally won the lottery.'[3] The key question, then, is whether the behaviors used in earlier experiments have a similar status.

To answer this question, we need to find a method that can detect behaviors that occupy this 'neither intentional nor unintentional' category. We can test to see that our method is working correctly by making sure that it accurately classifies the Lottery Vignette. Then we can see whether it obtains similar results for my original vignettes or whether it classifies those vignettes unambiguously as intentional or unintentional.

In search of such a method, I conducted a simple experiment. Subjects were 161 people spending time in a Manhattan public park. Each subject was randomly assigned to receive one of three stories — the Lottery Vignette, the Help Vignette, and the Harm Vignette. Within each of these conditions, subjects were randomly assigned to receive one of two possible questions. Some subjects were asked whether they agreed or disagreed with the claim that the behavior was performed *intentionally*; others were asked whether they agreed or disagreed with the claim that the behavior was performed *unintentionally*. Levels of agreement were assessed on a scale from -2 (disagree) to 2 (agree) with the 0 point marked "in between."

This new methodology makes it possible for subjects to classify behaviors into a broader array of categories. Subjects can say that it is right to call a behavior 'intentional,' that it is right to call the behavior 'unintentional,' or that it isn't right to call

[3] I am not quite sure why behaviors like this one do not fit comfortably into either category. Certainly, it is not a category mistake to apply the words 'intentional' and 'unintentional' to behaviors like winning the lottery. (We can easily imagine a person who wins the lottery unintentionally.)

To make matters even more confusing, the same phenomenon appears to arise for the sentences 'Jake intended to win the lottery' and 'Jake did not intend to win the lottery.'

the behavior either ‘intentional’ or ‘unintentional.’ The experimental results were as follows:

	Lottery	**Help**	**Harm**
Intentionally	-.5	-1.3	1
Unintentionally	-.5	1.5	-1.5

As expected, subjects classified the lottery behavior neither as intentional nor as unintentional. But they did not react in the same way to the original vignettes. Indeed, they seemed perfectly happy to say that the chairman harmed the environment intentionally and helped the environment unintentionally. These results appear to cast doubt on Meeks’s hypothesis.

But despite this apparent falsification, I feel certain that Meeks is on to something important. It really does sound wrong to ask a question like: ‘Did the chairman intentionally help the environment?’ For some reason, people seem reluctant to use the concepts *intentionally* and *unintentionally* in cases where there is no reason not to perform the behavior.

Perhaps this reluctance is due entirely to conversational pragmatics, but I suspect that there is something more afoot. Some preliminary evidence here can be derived from the study of other languages. In Russian, for example, there are two words that might be translated as ‘intentionally’ — *spetsalna* and *narochna*. The word ‘spetsalna’ can be applied to any behavior; ‘narochna’ can only be applied to behaviors that were in some way wrong or bad. If we now assume that any difference in pragmatics must involve

some difference in semantics, we must conclude that there is some semantic difference between these two words. We therefore seem forced to draw the conclusion that there is something in the very semantics of certain words that makes people reluctant to apply them in cases where there was no reason not to perform the behavior.

5. **Steven Sverdlik** focuses on the different attitudes one can have toward a foreseen side-effect. One can be *indifferent* to such a side-effect; one can bring it about *reluctantly*; or one can actively *seek to prevent* it. In a series of studies, Sverdlik examines the relationship between these various attitudes and people's use of the concept of intentional action. The results were quite surprising. Even when the effect itself was clearly bad, people only regarded it as intentional when the agent was indifferent, not when the agent was reluctant or trying to prevent it.[4]

These results have two key implications. One is fairly technical (and probably of interest only to researchers in action theory). The other is of broader interest, bearing on questions about the relationship between folk morality and folk psychology.

The technical implication has to do with the kinds of mental state ascriptions that figure in people's concept of intentional action. Previous researchers had usually defined the term 'side-effect' by saying that an outcome counts as a side-effect if (1) the agent did

[4] In one study, Sverdlik presented subjects with the case of a chairman who harms the environment and is indifferent to this environmental harm. The majority of subjects classified this behavior as unintentional. This result is surprising, since the behavior in question was classified as intentional by the majority of subjects in six independent studies (Knobe 2003a, 2004; Knobe & Burra forthcoming; McCann 2004; Malle 2004; Nichols 2004) and is therefore considered extremely robust.

Sverdlik informs me (personal communication) that his experiment was administered by two different people. He was able to look back through the original questionnaires and present separate data for each of the two experimenters. The results were telling. One experimenter obtained the usual pattern of results (with 64% of subjects saying that the agent acted intentionally); the other obtained a quite unusual pattern (with only 30% of subjects saying that the agent acted intentionally). The upshot was a statistically significant difference between the two experimenters, $\chi^2 (1, N=34) = 3.9, p < .05$. My guess is that the second experimenter did not simply hand out the questionnaires but also said something that biased subjects' responses.

not specifically want it to occur but (2) the agent did choose to perform a behavior that she knew would bring it about. In discussions of the concept of intentional action, most researchers had assumed that all side-effects could be treated in the same way. Sverdlik's results strongly suggest that this approach was mistaken. It seems that we also need to take account of a third feature, namely, (3) whether or not the agent specifically wanted the outcome *not* to occur. A full analysis of the folk concept of intentional action would describe the complex interplay of these three features in the process that generates people's intuitions.

The more philosophical implication has to do with the relationship between people's intentional action intuitions and their feelings of blame. Previous researchers had suggested that people's feelings of blame end up shaping their intuitions — the idea being that people consider certain side-effects to be intentional because they blame the agent for producing those side-effects (Adams & Steadman 2004; forthcoming; Malle forthcoming; Malle & Nelson 2003). Sverdlik's results call this view into question. When the agent reluctantly brought about a side-effect, people did not regard the side-effect as intentional *even though they felt that the agent was blameworthy*. It is hard to see how one might make sense of this result on the assumption that people's feelings of blame are what lead them to classify certain side-effects as intentional.

Perhaps the best way to explain this result would be to suppose that the process leading up to people's intentional action intuitions is a kind of heuristic. The process is an extremely quick and simple one, but it is constructed in such a way that it normally matches up with people's feelings of praise and blame. People do not normally praise or blame agents for reluctant side-effects, and the process is therefore constructed in such a

way that it classifies all reluctant side-effects as unintentional. In certain cases, however, reluctant side-effects may still elicit feelings of blame. In those cases, people's feelings of blame end up diverging in tell-tale ways from their intentional action intuitions.

6. By now, it should be clear that my critics have raised a wide array of important questions and that I have only been able to answer a small portion of them. Still, I take comfort from the fact that research on this issue has been progressing at an astonishing rate. It seems to me highly probable that other philosophers and psychologists will soon find a way to address the questions that have thus far escaped my grasp.

References

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

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

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*. *63,* 190-193.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*, 309-324.

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

Knobe, J. & Burra, A. (forthcoming). Intention and Intentional Action: A Cross-Cultural Study. *Journal of Culture and Cognition.*

Knobe, J. & Kelly, S. (2004). Can One Act for a Reason without Acting Intentionally? Unpublished manuscript. Princeton University.

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

Malle, B. F. (forthcoming). The Moral Dimension of People's Intentionality Judgments. *Journal of Culture and Cognition*.

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. (2004). More Evidence on Intentional Action and Intending. Unpublished manuscript. Texas A&M University.

Nadelhoffer, T. (forthcoming a). The Butler Problem Revisited. *Analysis*.

Nadelhoffer, T. (forthcoming b). Praise, Side Effects, and Folk Ascriptions of Intentional Action." Forthcoming in The Journal of Theoretical and Philosophical Psychology.

Nadelhoffer, T. (forthcoming c). Skill, Luck, Control, and Folk Ascriptions of Intentional Action. *Philosophical Psychology*.

Nichols, S. (2004). Unpublished data. University of Utah.

Sverdlik, S. (2004). Some Experiments Investigating the Commonsense Concepts of Moral Responsibility and Intentional Action. Unpublished manuscript. Southern Methodist University.