The Good, the Bad and the Blameworthy:
Understanding the Role of Evaluative Reasoning in Folk Psychology

Joshua Knobe Gabriel Mendlow
Princeton University

People ordinarily make sense of their own behavior and that of others by invoking concepts like belief, desire, and intention. Philosophers refer to this network of concepts and related principles as 'folk psychology.' The prevailing view of folk psychology among philosophers of mind and psychologists is that it is a proto-scientific theory whose function is to explain and predict behavior.

Recent studies call this view into question by suggesting that moral considerations play an essential role in the application of certain folk-psychological concepts (Knobe 2003a, 2003b, 2004; Knobe & Burra forthcoming; Nadelhoffer forthcoming). If the function of folk psychology were just to predict and explain behavior, moral considerations wouldn't have any obvious role to play. Since they do play a role, it seems probable that folk psychology has a function other than, or at least in addition to, that of predicting and explaining behavior. It is our assumption—an assumption our interlocutors presumably share—that, since moral considerations play an essential role in the application of certain folk-psychological concepts, getting clear on the nature and extent of this role is indispensable to arriving at a proper understanding of folk psychology.

Nadelhoffer's paper addresses the importance of moral considerations to the application of a specific folk-psychological concept—the concept of intentional action. He is concerned to explain, in particular, why people are more likely to regard a side-effect of an action as brought about intentionally when that side-effect is harmful than when it is beneficial. His hypothesis is that one's judgment as to whether a given side-effect was brought about intentionally is influenced by the amount of *praise* or *blame* one assigns. Since people typically blame the agent for bad side-effects but don't praise the agent for good ones, Nadelhoffer claims, they tend to regard bad side-effects as intentional and good ones as unintentional.

Here we propose an alternative hypothesis that in our view does a better job of making sense of the role played by the concept of intentional action in folk psychology and in people's lives. Our argument for the hypothesis draws on conceptual considerations as well as new empirical data.

I

Before presenting the alternative hypothesis, we'd like to dwell for a moment on a distinction vital to our interpretation of the phenomena: the distinction between praiseworthiness and blameworthiness (on the one hand) and goodness and badness (on the other). A few thought experiments should suffice to show that these two pairs of concepts are different. First, let's consider some cases in which actions are bad without being blameworthy or good without being praiseworthy. Suppose that while mowing the lawn you unwittingly destroy the last specimen of a species of mushroom. There is clearly a respect in which your action was bad; extinguishing a species is by any account a bad thing. But if you had no reason to believe that the lawn was home to an endangered species, it is plausible to think that your action was not worthy of blame. Changing the example somewhat, suppose that while watering the lawn you unknowingly save that same species of mushroom from extinction. No doubt your action was good—yet it can hardly be said to be worthy of praise, as you had no intention of saving the mushroom. Second, consider cases in which actions are blameworthy but not bad or praiseworthy but not good. If an act of attempted murder has no harmful consequences, it is blameworthy without being bad. If an act of attempted heroism has no beneficial consequences, it is praiseworthy without being good.

Furthermore, there appear to be actions that are simultaneously praiseworthy and bad or blameworthy and good. Consider, as an instance of the latter conjunction, the act of pushing an old lady out of the path of a truck. Pushing a non-consenting person is, taken in itself, a bad thing. But it is clearly praiseworthy under the circumstances. Note that while the goodness of the consequence—saving the woman's life—outweighs the badness of the act itself, the goodness of the consequence does not *cancel* the badness of the act. Despite its beneficial consequence and resultant praiseworthiness, the act of

pushing an old lady remains a bad thing. That's why it makes sense to say something along the lines of "It's *too bad* you had to push that old lady in order to save her." That an action can be simultaneously good and blameworthy is no more difficult to show. Consider a pedophile who comforts a lonely boy in order to win his trust. The act of comforting a child is, taken in itself, a good thing, yet the pedophile's sinister motive renders the action blameworthy. That the blameworthiness of the action is independent of its possible consequences is evidenced by the fact that one would still ascribe blame to the pedophile even if no subsequent abuse took place. Thus an action can be blameworthy yet good in itself.

Given that praiseworthiness can diverge from goodness and blameworthiness from badness, it is important that one be sensitive to the aforesaid distinction when explaining why moral considerations sometimes influence people's attributions of folk-psychological concepts. We believe, with respect to the present case, that the moral considerations that influence people's judgments as to whether a side-effect was intentional have to do, in the first instance, with the goodness or badness of the side-effect rather than with its praiseworthiness or blameworthiness.

II

Nadelhoffer apparently disagrees, suggesting that the results be understood in terms of the degree of praise or blame people assign to the agent. In general, he points out, people blame the agent for bad side-effects but do not praise the agent for good side-effects. Nadelhoffer suggests that this asymmetry in the assignment of praise and blame explains the asymmetry in people's application of the concept of intentional action. His hypothesis is that people regard the bad side-effect as intentional only because they blame the agent for it. If they had felt that the agent did not deserve blame, they would not have concluded that he brought about the side-effect intentionally.

From this hypothesis Nadelhoffer derives a new prediction. The hypothesis, recall, is that, under typical circumstances, good side-effects are not regarded as intentional because the agent is not given praise for bringing them about. Nadelhoffer predicts on the basis of this hypothesis that if an agent actually were regarded as

praiseworthy for bringing about a particular side-effect, people would be inclined to say that the agent brought about that side-effect intentionally.

In support of this prediction, Nadelhoffer reports new experimental results. Subjects were given a story about an agent who decides to help his friend even though he knows that doing so will decrease his own chances of winning a contest. The circumstance is one in which the agent seems praiseworthy for the bringing about of a side-effect, namely, a decrease in his chances of winning a contest.[1] As predicted, a substantial portion of subjects judged that the agent intentionally decreased his chances of winning. Nadelhoffer takes this result as evidence for the view that judgments of praise and blame influence whether side-effects are regarded as intentional.

We want to propose an alternative hypothesis: judgments of praise and blame have little or no impact on whether side-effects are regarded as intentional; the observed effects are due instead to the perceived goodness or badness of the side-effect itself. If this hypothesis is correct, the results from Nadelhoffer's experiment cannot be explained in terms of the praise people accord the agent. We suggest that people's intuitions are influenced by a belief that decreasing one's chances of winning is, taken in itself, a bad thing. Of course the side-effect isn't *morally* bad, and it is presumably not bad on the whole, given its beneficial consequences for the agent's friend. Yet there is a clear sense in which decreasing one's own chances of winning can be classified as bad. As with the case of pushing the old lady, it is appropriate to say something along the lines of "It's *too bad* you had to sacrifice your own chances in order to help your friend." Such a remark would make sense only if the action were bad in some respect. Our hypothesis is that the intrinsic (if perhaps outweighed) badness of the side-effect itself is what influences people's intuitions about whether it was intentional.

As far as we can see, our alternative hypothesis accounts for the experimental data just as well as Nadelhoffer's does—no worse and no better. Thus, the experimental data reported here cannot help us decide between the two hypotheses. To find considerations that favor one hypothesis over the other, we need to turn elsewhere.

---

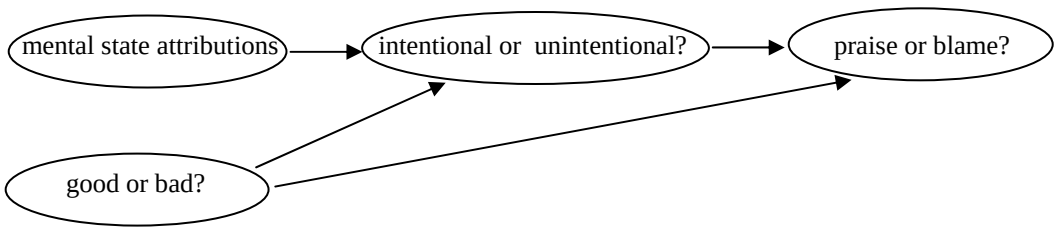
[1] We suspect, however, that people would be uncomfortable saying that the agent deserves praise for the side-effect as such. There is something awkward about saying that the agent is praiseworthy *for decreasing his chances of winning*. It would be more natural to say that the agent deserves praise *for helping his friend*.

II

We now consider two arguments for the view that blame has little direct influence on people's intuitions about whether a side-effect was brought about intentionally. One argument is *a priori* and relies on assumptions about the role that the concept of intentional action plays in people's lives. The other is empirical and draws on people's intuitions about particular cases.

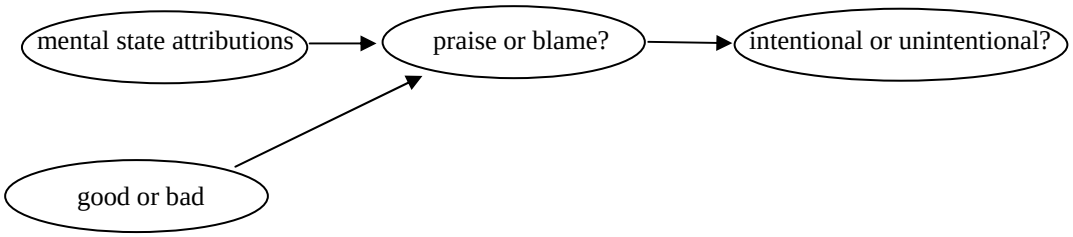
First, the *a priori* argument. Here we will be relying on the assumption that people's concept of intentional action is in some way *useful*. While further investigation might reveal the concept to be partially or entirely pointless, it seems reasonable to start from the assumption that the concept plays some helpful role in people's lives. Armed with this assumption, we can put two competing models of the concept to the test.

First consider a model that looks like this:



Here one's perception of the goodness or badness of a behavior influences one's judgment as to whether it was intentional, which in turn informs one's assignment of praise or blame. On this model, it is easy to see how the concept of intentional action plays a useful role: the model is consonant with the widely-held view that the concept of intentional action plays a role in the process by which people arrive at judgments of praise and blame.

Now consider a slightly different model:



On this second model, one's perception of the goodness or badness of a behavior directly influences one's assignment of praise or blame, which in turn informs one's judgment as to whether the behavior was intentional.

This second model is not compatible with the commonsense view that people invoke the concept of intentional action when they are determining whether or not to assign praise or blame. Instead, the model has people deciding whether or not to assign praise or blame *before* they have determined whether or not the behavior was performed intentionally. Thus this model attributes to the folk psychologist a seemingly *pointless* mechanism—one that serves neither the purpose of prediction/explanation nor that of moral judgment.

While it is in principle possible that folk psychology includes a pointless mechanism, the considerations adduced here provide preliminary support for the first model over the second.

That said, we turn to the empirical argument. This argument draws on people's intuitions about particular cases. In cases where people both regard a side-effect as bad and blame the agent for bringing it about, the two models make identical predictions. But in cases where people regard the side-effect as bad but do *not* blame the agent for bringing it about, the predictions diverge. If judgments of praise and blame are directly influencing people's application of the concept of intentional action, people should be overwhelmingly inclined to see these side-effects as unintentional. But if people's judgments about the goodness or badness of the behavior itself directly influence their application of the concept of intentional action, one would expect to find results much

like those obtained in Nadelhoffer's own experiment—with a substantial portion of subjects saying that the agent brought about the side-effects intentionally.

Let us turn, then, to a specific case, which we shall call the *Sales 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."

Consider the status of the side-effect *decreasing sales in New Jersey*. It seems clear that this side-effect is in some sense a bad one. But since Susan was actually increasing sales on the whole, people would probably be reluctant to say that she was worthy of blame. Thus, our two models lead to two different predictions. If blame is what matters, one would expect most people to say that Susan decreased sales unintentionally. But if the badness of the behavior itself is what matters, one would expect a substantial portion of the people to say that Susan decreased sales intentionally.

To decide between these competing hypotheses, we ran a simple mini-study. Subjects were 20 people spending time in a Manhattan public park. All subjects received a questionnaire containing the Sales Vignette followed by two questions. The first question was: 'Did Susan intentionally decrease sales in New Jersey?' The second question was: 'Does Susan deserve any praise or blame for decreasing sales in New

Jersey?' Subjects answered this second question by choosing between three options: (1) 'Susan deserves *praise* for decreasing sales in New Jersey.' (2) 'Susan deserves *blame* for decreasing sales in New Jersey.' (3) 'Susan deserves *neither praise nor blame* for decreasing sales in New Jersey.'

The percentage of subjects giving each combination of answers is displayed in the table below.

	Praise	**Blame**	**Neither**	*Total*
Intentionally	10%	5%	60%	75%
Unintentionally	0%	5%	20%	25%
Total	10%	10%	80%	

The key results here are straightforward. The vast majority of subjects said that the agent performed the behavior intentionally (75%) even though subjects also said that she deserved neither praise nor blame for performing it (80%). Most importantly, of those subjects who said that the agent deserved neither praise nor blame, a clear majority (75%) said that the agent performed the behavior intentionally.

This result spells trouble for any view according to which praise and blame are influencing people's intuitions as to whether or not a behavior was performed intentionally. Indeed, we see no plausible way to reconcile this result with the hypothesis that people regard side-effects as intentional only when they also regard the agent as praiseworthy or blameworthy.

III

We began by distinguishing judgments that an agent is praiseworthy or blameworthy from judgments that a behavior is good or bad. Our inquiry was concerned to determine which of these two kinds of judgment influences people's application of the concept of intentional action. The available evidence seems to indicate that people's application of the concept is influenced by judgments of goodness and badness without the mediation of judgments of praise and blame.

Works Cited

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. and A. Burra. (forthcoming). What is the Relation between Intention and Intentional Action. *Journal of Cognition and Culture.*

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