Intentional Action in Folk Psychology: An Experimental Investigation

Joshua Knobe

Princeton University

Forthcoming in *Philosophical Psychology*

Abstract: Four experiments examined people's folk-psychological concept of intentional action. The chief question was whether or not *evaluative* considerations — considerations of good and bad, right and wrong, praise and blame — played any role in that concept. The results indicated that the moral qualities of a behavior strongly influence people's judgements as to whether or not that behavior should be considered 'intentional.' After eliminating a number of alternative explanations, the author concludes that this effect is best explained by the hypothesis that evaluative considerations do play some role in people's concept of intentional action.

Intentional Action in Folk Psychology:

An Experimental Investigation

People normally draw a distinction between behaviors that are performed *intentionally* and those that are performed *unintentionally*. It can hardly be denied that this distinction occupies an important place in folk psychology, but researchers disagree about precisely how the distinction should be understood and what function it serves in people's lives. Indeed, looking through the existing literature on the concept of intentional action, one can discern two fundamentally divergent viewpoints.

On the first of these viewpoints, people's concept of intentional action is understood as one element in a tacit 'theory of mind' (Astington 1999, 2001), which is then understood as something like a scientific theory of human behavior (Gopnik & Meltzoff 1997; Gopnik & Wellman 1992). The basic intuition here is that, by classifying behaviors as intentional or unintentional, people are making a distinction that helps them to predict and explain behavior. This predictive and explanatory role is then held to be the key to understanding the nature of the concept itself. Of course, adherents of this viewpoint do not deny that the concept of intentional action is also used in various other kinds of reasoning (e.g., in reasoning about moral praise and moral blame), but these other uses are regarded as parasitic or secondary, not fundamental to the nature of the concept as such (Mele & Sverdlik 1996; cf. Churchland 1981, 1991).

I am grateful for comments from Alison Gopnik, Gilbert Harman, Bertram Malle, Michael Morris, Alfred Mele and Daniel Rothschild.

According to the second viewpoint, by contrast, people's concept of intentional action is bound up in a fundamental way with *evaluative* questions — with questions about good and bad, right and wrong, praise and blame. Adherents of this second viewpoint claim that people's concept of intentional action can only be correctly understood when we see that it is used not only to predict and explain behaviors but also to determine the moral significance of those behaviors (Bratman 1987). Taking this second viewpoint to a more radical extreme, a number of researchers have argued that people actually use their moral beliefs when they are trying to determine whether or not a given behavior is intentional (Harman 1976; Lowe 1978; Pitcher 1970).

Although these two viewpoints have been discussed primarily by philosophers, it seems clear that many of the crucial issues in the debate can be illuminated by systematic psychological experiments, and that is the approach that I will be adopting here. By studying the precise conditions under which people consider behaviors to be 'intentional,' I provide support for an empirical hypothesis about people's concept of intentional action. This hypothesis is then shown to have implications for the broader questions about the function that the concept of intentional action serves in people's lives.

Intentional Action and Skill

The hypothesis that I will be defending here involves a substantial a revision in an account that the psychologist Bertram Malle and I put forth a number of years ago (Malle & Knobe 1997a). I therefore begin by briefly reviewing one aspect of that earlier account.

Malle and I (1997a) set out to understand people's concept of intentional action. How do people determine whether a given behavior is 'intentional' or 'unintentional'?

We began by simply asking subjects what it meant for a behavior to be 'intentional.' Subjects responded by describing the mental states that, they believed, accompany all intentional action. For example: 'The person meant to act that way and was motivated to do so.' Or: 'Someone gave thought to the action beforehand and chose to do it.'

It seemed to us, however, that these mental states alone were not sufficient to make people consider a behavior intentional. Suppose that Bob is trying to win the lottery and actually does win the lottery. One would not normally say that such a person 'intentionally' won the lottery. The problem here is that Bob doesn't have the requisite sort of control over his performance. There is a kind of match between what he was trying to do and what he actually ended up doing, but this match seems to be too much of a coincidence. He doesn't have any reliable mechanism by which to transform his attempt into the corresponding behavior.

Of course, one could imagine that various other mental states were added in to this story. One could suppose that the agent *believed* that he would win or even that he *intended* to win. But no matter what mental state you add, people still won't call the behavior 'intentional,' because the issue here has nothing to do with mental states: it has to do with the amount of control that the agent actually has over the outcome. Or, to put it in the terms of our (1997a) paper, it appears that *skill* is one component of people's concept of intentional action.

In a series of studies, we asked subjects to read vignettes about an agent who tries to perform a behavior and actually does perform that behavior but lacks the skill to perform the behavior reliably. Most subjects said that the agent was *trying* to perform the behavior but was not performing that behavior *intentionally*. These results seemed to lend support to our hypothesis.

Skill and Evaluative Considerations

But here one may begin to wonder *why* the concept of intentional action might include a skill component. The concept of intentional action can be seen as a kind of tool, used in the performance of certain tasks. The question, then, is: Do people actually do a better job of performing these tasks because the concept of intentional action includes a skill component, or would they be able to perform those tasks just as well if the concept of intentional action had been slightly different?

When we think of the issue in this light, we may be struck by the fact that people often use the concept of intentional action when they are assigning praise and blame. It therefore seems natural to consider the possibility that skill plays a role in people's concept of intentional action because it, too, is intimately related to the assignment of praise and blame. Suppose that Kate gives the correct answer to a difficult math problem. We would give her more praise if she actually had the ability to reliably give correct answers to such problems than we would if she just got the right answer by sheer luck. And perhaps that is the reason why, if her success were due to luck alone, we would be reluctant to say that she *intentionally* gave the right answer. In other words, one might conjecture that the skill component is a useful aspect of our concept of intentional action because it helps us to distinguish between behaviors that deserve full praise or blame and behaviors that only deserve a diminished amount of praise or blame (perhaps even no praise or blame at all).

But now we face a puzzle. It seems as though there is a very complex relationship between judgements of skill and judgements of praise and blame. It's not as though people always assign greater amounts of praise or blame when they believe that the agent had greater skill. Depending on the particular behavior, the particular circumstances, etc., praise and blame

might be affected by skill in any number of different ways. So it may appear that one ought to distinguish between a number of different cases — those in which the agent performs a difficult behavior and would be praised more if she had the requisite skill, those in which the agent performs an immoral behavior and would receive approximately the same amount of blame whether she had skill or not, and so on.

When Malle and I were designing our experiments, we didn't distinguish between these various cases. So, for example, when we were trying to show that people don't call a behavior 'intentional' if the agent lacks skill, we didn't use any vignettes in which there seemed to be a possibility of moral blame. But we didn't think that would matter. Our assumption was that either skill plays a role in the concept of intentional action, or it doesn't. If it plays a role in cases where the agent is trying to win a game of darts, then it will also play a role in cases where the agent is trying to perform some immoral behavior. Our unstated assumption, in other words, was that the concept would be *independent of evaluative considerations*.

I was woken from this dogmatic slumber by Mele's (2001) rigorous and thought-provoking response to our work. Mele discusses each of our claims in detail, pointing to a number of important areas in which he believes that our theory needs revision. But when he turns to the proposed 'skill component,' he argues that our theory is correct as it stands. He argues that people's concept of intentional action includes skill as a necessary condition (cf. Mele & Moser 1994) and that the concept is entirely independent of evaluative considerations (cf. Mele & Sverdlik 1996). But Mele goes beyond us in a number of important respects. First, he takes our unstated assumption and frames it instead as an explicit hypothesis. Second, he proposes an ingenious series of experiments to prove that this hypothesis is correct.

No psychology experiment can ever be completely decisive. No matter how well-designed an experiment is, it will almost certainly be possible to construct alternative explanations of the results. Still, I am inclined to think that Mele's (2001) proposed experiments provide a fairly good test of his hypothesis. The only problem is that when these experiments are actually performed, they don't come out the way he thought they would. On the contrary, all of the results point to the conclusion that evaluative considerations actually do play an important role in people's concept of intentional action.

Experiment 1: Intentional Action and Moral Blame

We claimed above that people would not normally say that Bob won the lottery *intentionally* even if he was trying to win the lottery and actually did win. But suppose we consider a case in which there is more potential for moral blame. Janet is trying to kill the president. It is very unlikely that she will succeed. In fact, let us suppose that she is no more likely to succeed at killing the president than Bob was at winning the lottery. Nonetheless, she happens to be incredibly lucky, and she actually does manage to carry out her plan. In such a case, we would surely say that she had *intentionally* killed the president. But this seems puzzling, since her chance of success was, by hypothesis, no greater than Bob's chance of winning the lottery. Might it turn out that our reasoning here is influenced by evaluative considerations?

Of course, there are a great number of differences between trying to win the lottery and trying to kill the president, and any one of these differences could account for the difference in our intuitions about the two cases. Fortunately, Mele suggests a more closely parallel series of cases. Consider a person who fires a gun at a target and actually hits that target. And now consider two ways in which the case can vary. First, the target can be either (1) something that it

would be an achievement to shoot or (2) something that it would be morally wrong to shoot. Second, the agent can be either (a) a skilled marksman who easily hits the target or (b) an inexperienced marksman who just happens to get lucky.

Crossing these two variables, we get four possible options, which I spell out in more detail:

(1A) *Achievement/Skill*: 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.

Jake is an expert marksman. His hands are steady. The gun is aimed perfectly…

The bullet lands directly on the bull's-eye. Jake wins the contest.

(1B) *Achievement/No-Skill*: 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.

(2A) *Immoral/Skill*: 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.

Jake is an expert marksman. His hands are steady. The gun is aimed perfectly…

The bullet hits her directly in the heart. She dies instantly.

(2B) *Immoral/No-Skill*: 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.

Notice that the 'no-skill' conditions lie somewhere in the gray area between total skill and total luck. In these conditions, the agent's success does not depend almost entirely on skill (as it does in the 'skill' conditions), nor does the agent's success depend almost entirely on luck (as it would, e.g., in a lottery). Though the agent benefits quite a bit from luck, he does have some amount of control over the outcome.

The key question is whether people's judgements about cases like these will depend in part on the moral qualities of the behavior. Are people more likely to say that the agent acted intentionally when he is trying to perform the immoral behavior than they are when he is trying to perform the achievement behavior?

To address this issue, I conducted a simple experiment. Subjects were 80 people spending time in a Manhattan public park. Each subject was presented with one of the four vignettes presented above (1A, 1B, 2A, or 2B). Each subject then received two questions: ‘Was Jake *trying* to kill his aunt [to hit the bull’s-eye]?’ and ‘Did Jake *intentionally* kill his aunt [hit the bull’s-eye]?’

The question about ‘trying’ was included as a test that subjects understood the story, and indeed 91% of subjects said that Jake was trying. The key dependent variable was subjects’ answers to the question about whether or not the agent was acting ‘intentionally.’

Here are the percentages of subjects who said that the agent was acting ‘intentionally’ within each condition:

	Immoral	Achievement
Skill	95%	79%
No Skill	76%	28%

From these results, it certainly appears that evaluative considerations have some impact on people’s judgements. Notice in particular the results for vignettes in the no-skill conditions. Only 28% of subjects in the achievement condition said the behavior was intentional, but a full 76% of subjects in the immoral condition thought the behavior was intentional. This contrast is statistically significant, $\chi^2 (1, N=37) = 7.7, p < .01$.

Interestingly enough, these results were predicted by Mele himself. Mele does not deny that people are sometimes more likely to apply the word ‘intentional’ to behaviors that they perceive to be morally blameworthy. So he fully admits that evaluative considerations have an impact on people’s use of the *word* ‘intentional.’ His claim is simply that this impact should not be ascribed to people’s *concept* of intentional action.

Two Ways of Thinking about Concepts

Here we need to distinguish two different ways of thinking about the relation between concepts and people’s use of words in ordinary language.

On one view, a concept has something to do with the *correct* use of a word. When we offer an account of, e.g., the concept of intentional action, we are offering a standard against which people's ordinary utterances can be judged. If people don't actually use the words 'intentional action' in the way specified by the account, we might conclude that they are speaking incorrectly. This may be a legitimate way of thinking about concepts, but it certainly isn't the type of inquiry under discussion here.

When Malle and I set out to understand the 'folk concept' of intentional action, our aim was to investigate people's actual understanding of intentional action. In an inquiry of this latter type, the *concept* of intentional action should be understood as a psychological state posited by a scientific theory. Hypotheses about this state can be tested in a fairly straightforward way using observation and experiment. In essence, we test hypotheses about people's concept of intentional action by examining the behaviors that, we believe, are influenced by this concept.

Presumably, people's concept of intentional action influences their use of the word 'intentional.' But it may also influence other behaviors that they perform — behaviors that have nothing to do with using the word 'intentional.' As we said above, people's concept of intentional action influences their decisions about when to offer praise or blame (Shaver 1985; Young 2001). It also influences their explanations of other people's behaviors (Malle, Knobe, O'Laughlin, Pearce & Nelson 2000) and their decisions about which behaviors to explain (Malle & Knobe 1997b). By looking carefully at all of these phenomena, we can gain valuable evidence about the concept itself.

When we are thinking about concepts in this latter sense, it makes no sense to say that people have somehow failed to correctly grasp the concept of intentional action. The concept just

is whatever people have grasped. So if people's concept of intentional action isn't what our theory says it is, then it is our theory, not the concept itself, that has to change.

Still, it is essential to distinguish between people's concept of intentional action and the conditions under which people actually use the word 'intentional.' People's use of the word 'intentional' can be influenced by all sorts of factors. For example, it could be influenced by the state of the person's tongue. (If a person's tongue is injured, he or she might be less likely to use the word 'intentional.') It might also be influenced by the person's goals in the conversation, by his beliefs about the specific situation in question, and so forth.*

Mele's point (if I understand him correctly) is that, even if it turns out that people are more likely to call an action 'intentional' when that action is immoral, we shouldn't necessarily ascribe this effect to their concept of intentional action. We should ascribe it instead to some other factor. He then proposes two experiments which, he believes, will show that the effect is actually due to one of these other factors.

Experiment 2: Expressing Moral Disapproval

Mele's first suggestion is that we give people an opportunity to offer two separate kinds of judgements — a judgement as to whether the action is *intentional* and then, separately, a judgement as to whether the action is *blameworthy*. His prediction is that, if people are permitted

---

* Here we face an intriguing (and thus far unresolved) issue: Why is it that different subjects in experiment 1 gave different responses to the same question? Is it because different people actually have different concepts of intentional action? Or is it because, although everyone shares the same concept of intentional action, people differ with respect to other mental states that influence their use of word 'intentional'? I am now in the process of designing experiments to resolve this issue.

to offer these two types of judgements, they will tend to say that immoral behaviors performed without skill are unintentional.

To explain the reasoning here, it may be helpful to introduce an analogy. Suppose that we were trying to understand people's concept of *rape*. We could give them short vignettes and then ask them whether the actions described in these vignettes count as 'rape.' But responses to this question might not accurately reflect people's concept. Suppose, e.g., that we present subjects with the story of a man who uses particularly despicable forms of verbal pressure to make his girlfriend have sex with him. If subjects simply said 'That's not rape,' they might appear to be condoning the man's actions. So they may label the action 'rape' because that's the only means they have available of expressing their disapproval. The simplest solution to this problem would be to ask subjects not only whether they thought the action was rape but also whether they thought the action was morally blameworthy. That way, subjects would have available the option of saying that the action was definitely immoral but that it nonetheless didn't count as 'rape.'

Mele suggests that there may be a similar problem with the method we used to get at people's concept of intentional action. After all, we simply present subjects with a behavior and ask them whether that behavior counts as 'intentional.' We don't give them any way of further clarifying their attitude toward the behavior. But if they say that the behavior isn't intentional, it may appear that they are excusing it. That is, it may appear that they are saying something like: 'Oh, that's just an unintentional behavior — an accident — and the agent shouldn't be blamed for it.' The solution, here as in the analogous case, is to give subjects an independent opportunity to express their moral disapproval.

To address this issue, I conducted a second experiment. Subjects were 68 people spending time in a Manhattan public park. This time, there were only two conditions. Subjects in

the ‘achievement’ condition received the achievement/no-skill questionnaire used in experiment 1, along with the additional item: ‘How much praise does Jake deserve for what he did?’ (This additional item preceded the questions about whether Jake was trying and whether he acted intentionally.) Similarly, subjects in the ‘immoral’ condition received the immoral/no-skill questionnaire used in experiment 1, along with the additional item: ‘How much blame does Jake deserve for what he did?’ Subjects indicated the appropriate amount of praise or blame on a scale from 0 [no praise or blame] to 6 [a lot of praise or blame].

Each subject also received a second questionnaire. Subjects in the achievement condition received the achievement/skill vignette from experiment 1 followed by the question: ‘In this second story, how much praise does Jake deserve for what he did?’ Subjects in the immoral condition received the immoral/skill vignette from experiment 1 followed by the question: ‘In this second story, how much blame does Jake deserve for what he did?’ This second questionnaire made possible an additional analysis, which I report in a later section of the present paper.

Overall, the results showed that giving subjects a chance to express praise or blame in no way diminished the asymmetry found in experiment 1. Once again, almost all subjects (96%) said that Jake was ‘trying.’ And, once again, the blameworthiness of the behavior had a major impact on subjects’ judgements as to whether or not the behavior was ‘intentional.’ When the behavior was presented as an achievement, only 23% of subjects thought that it should be called ‘intentional’ if the agent lacked skill. But when the behavior seemed immoral, a full 91% thought the action could be called ‘intentional’ even if the agent didn’t have skill. This difference is highly significant, χ^2 (1, N=68) = 31.9, $p < .001$.

In short, people claimed that the immoral/no-skill behavior was intentional even though they quite clearly had the option of saying that the action was unintentional but nonetheless blameworthy. This result suggests that subjects genuinely believed the immoral/no-skill behavior to be intentional and were not simply calling it intentional as a way of indicating that it was deserving of blame.

Experiment 3: Theoretical Beliefs

Mele's second argument is that people's judgements may be influenced by fairly abstract beliefs about the nature of moral blameworthiness. In particular, people may judge certain behaviors to be 'intentional' because they believe that only intentional behaviors can be morally blameworthy.

To see the force of this objection, it may be helpful to consider another analogy. Suppose that we were studying a given person's grammatical intuitions. When we present him with a sentence, it immediately appears to him to be grammatical, ungrammatical or somewhere in between. (Presumably, he has no idea what rules he is using to make this judgement.) Our task is to figure out what intuitions he has. How will we proceed?

The obvious approach would be simply to ask him questions like 'Is this sentence grammatical?' — and this approach will usually be successful. But there is always a danger. Suppose that the person has learned certain 'grammatical rules' in high school English classes. These rules may not fit perfectly with his grammatical intuitions. So when we present him with a sentence, he might think: 'Well, that looks grammatical' — and then: 'But wait! It violates the rules I learned in high school. So it must not be grammatical after all.'

Mele thinks that something similar might be going on when people make judgements about whether or not a given behavior is intentional. They may have some theoretical beliefs about the relation between moral blameworthiness and intentional action. Specifically, they may believe that no behavior can ever be morally blameworthy if it is unintentional. Then, when they are confronted with a morally blameworthy behavior that seems unintentional, they may think, in effect: 'But wait! Since this behavior is blameworthy, it just *couldn't* be unintentional.'

Perhaps Mele is right in thinking that some of our subjects have fairly theoretical beliefs about intentional action and moral blame, but I don't think that these theoretical beliefs play any major role in their judgement that the immoral/no-skill behavior is intentional. In other words, it might turn out that some people believe that all blameworthy behaviors are intentional, but even if they didn't have this moral belief, I think they would still say that blameworthy behaviors performed without skill can be intentional.

To test this claim, we need to introduce a manipulation that causes subjects to believe that actions can be blameworthy even if they aren't intentional and then try to figure out whether or not these subjects still say that the immoral/no-skill behavior is intentional. Mele himself suggests a good way of achieving this effect. We can first expose subjects to a vignette about a person who performs a behavior that is unintentional but apparently blameworthy. He suggests the following story:

> Bob got rip-roaring drunk at a party after work. When the party ended, he stumbled to his car and started driving home.
>
> He was very drunk at the time — so drunk that he eventually lost control of his car, swerved into oncoming traffic, and killed a family of five. (Mele 2001, p. 41)

He then predicts that subjects exposed to this vignette will offer different responses to the vignettes we used in experiment 1. Once they see that a behavior can be unintentional even if it is blameworthy, he claims, they should be willing to say that the immoral/no-skill behavior, though highly blameworthy, is nonetheless unintentional.

I performed an experiment to test this hypothesis. Subjects were 61 people spending time in a Manhattan public park. Each subject was presented with a two-page questionnaire. On the first page was the vignette about the drunk driver. On the second page were precisely the same vignettes used in the no-skill conditions of the previous study.

Four subjects failed to answer at least one question on the questionnaire. Analyses were performed on the remaining 57 subjects.

As predicted, most subjects judged that the drunk driver's behavior was unintentional but blameworthy nonetheless. A full 96% said that the behavior was unintentional. When subjects were asked to rate the driver's blameworthiness on a scale from 0 (no blame) to 6 (a lot of blame), the mean rating was 5.3 (almost the maximum possible blame).

Nonetheless, judgements about the vignettes on the second page showed the familiar asymmetry: 84% said that the immoral behavior was intentional, whereas only 40% said that the achievement behavior was intentional. This contrast is highly significant, χ^2 (1, N=57) = 12.1, p < .001.

How is this asymmetry to be explained? It can hardly be claimed that subjects held some theoretical belief to the effect that all blameworthy behaviors are intentional. After all, they had just declared that the drunk driver's behavior, though highly blameworthy, was perfectly unintentional. The most plausible explanation is that subjects were guided, not by any theoretical beliefs about the nature of morality and intentional action, but rather by the tacit understanding

they normally use when deciding whether or not certain actions are intentional, i.e., by their concept of intentional action.

Towards a New Hypothesis

In our earlier theory (Malle and Knobe 1997a), we described skill as a 'necessary condition' in people's concept of intentional action. This earlier theory appears to be false. A correct theory, it seems, would assign a far more complex role to people's beliefs about skill. Specifically, it would explain the ways in which beliefs about skill interact with judgements of praise and blame.

Notice, first of all, that people sometimes give the agent less credit or blame for performing a behavior when they believe that the agent didn't really have the skill to perform that behavior reliably. Thus, people may give the agent considerably less praise for an achievement if they ascribe that achievement primarily to luck. In such a case, people may regard the behavior itself as meritorious but they nonetheless choose not to praise the agent for performing that behavior. It is as though a certain sort of evaluative link between the behavior and the agent has been severed. Let us say, in such cases, that people *dissociate* the agent from the behavior.

In other words, we will say that people dissociate the agent from the behavior when they conclude that the agent deserves less praise or blame because of her lack of skill. To get a rough sense for the degree to which a person has dissociated the agent from a behavior, we can first ask the person how much praise or blame the agent deserves for the behavior, then ask the person how much praise or blame the agent would have deserved if she had had enough skill to perform the behavior reliably, and finally check to see how great a difference there is between these two

judgements. Our general rule will be: the greater the difference, the greater the degree of dissociation.

But, of course, we are not claiming that dissociation just *is* the difference between these two judgements. Rather, dissociation is a psychological state whose existence can be *inferred* from a difference between two judgements. The idea is that, in the process by which people determine how much praise or blame the agent deserves, they sometimes arrive at a psychological state whose content is roughly: 'The agent deserves less praise or blame for this behavior because of her lack of skill.' It is this psychological state that we are referring to with the word 'dissociation.'

We can now introduce a new hypothesis. Perhaps what matters is not the degree to which the agent is perceived to lack skill but rather the degree to which the agent is dissociated from the behavior. Judgements of skill serve as an input to the process by which people determine the degree of dissociation. And the degree of dissociation is then an input to people's concept of intentional action. The more the agent is dissociated from his action, the more reluctant people will be to consider that action intentional. Clearly, the experiments presented thus far are not by themselves sufficient to confirm this hypothesis. Still, we can look to our results for some preliminary support.

In experiment 2, each subject gave ratings of praise or blame for both the 'skill' and 'no-skill' versions of a single vignette. By comparing the ratings that a given subject gave in these two versions, we can get an approximate sense of the degree to which that subject dissociated the agent from his action. For example, suppose that a subject gives the agent a praise rating of 5 for hitting the bull's-eye by skill but only gave the agent a praise rating of 3 for hitting the bull's-eye

by luck. It appears, then, that the subject is decreasing his praise rating by 2 points in the no-skill version of the vignette. Let us say, then, that this subject's 'dissociation score' is 2.

In general, subjects showed higher dissociation scores when presented with the achievement condition than they did when presented with the immoral condition. On average, subjects gave the agent 1.7 points less praise when he only managed to hit the bull's-eye by luck than they did when he hit the bull's-eye by skill. By contrast, subjects said that the agent deserved .3 points less blame when he only managed to kill his aunt by luck than when he killed her by skill. This contrast is statistically significant, t (66) = 3.12, $p < .01$.

More importantly, there was a significant correlation such that subjects with high dissociation scores were more likely to say that the no-skill behavior was unintentional, r (68) = -.36, $p < .01$. Although this result may be susceptible to a number of alternative explanations, it does lend at least preliminary support to the present hypothesis.

Experiment 4: Dissociation vs. Blame

On the hypothesis we have just put forth, people's judgements of praise and blame stand in a very complex relation to their judgements about whether or not a behavior is intentional. The critical factor turns out to be, not the praise or blame that people actually assign to the behavior, but rather the *degree of dissociation* — which is the degree to which people think that the agent deserves reduced praise or blame because of her lack of skill. The hypothesis says that people don't regard the behavior as intentional when there is a high degree of dissociation.

But here we should consider a rival hypothesis. Perhaps what matters is simply the amount of blame that people assign to the behavior. For example, it may be that there are two processes at work: (a) people have a general tendency to regard behaviors as unintentional when

the agent lacks skill, but (b) when people feel that the behavior is especially blameworthy, they override this general tendency and consider the behavior intentional even when the agent lacks skill. This basic position has been advanced by Harman (1976), by Lowe (1978), and, in a somewhat different form, by Pitcher (1970). I will refer to it as the 'blame hypothesis.'

Although I have been using the notion of dissociation to explain the results obtained thus far, it is clear that all of these results could equally well be explained in terms of blame. For example, suppose that we reanalyze the data from experiment 2 with the blame hypothesis in mind. We can assign each subject a 'valence score' by counting praise ratings as positive numbers and blame ratings as negative numbers. This valence score then turns out to be correlated with people's judgements about whether the behavior is intentional, $r\,(66) = -.66$, $p < .001$, indicating that the blame hypothesis is perfectly capable of explaining the pattern of results.

To truly discriminate between the blame hypothesis and the dissociation hypothesis, we need to present subjects with a case in which the two hypotheses generate different predictions. I therefore ran an experiment in which subjects were presented with a vignette about a *morally good* behavior.

Subjects were 36 people spending time in a Manhattan public park. All subjects received the same questionnaire. The first page contained the 'no-skill' version of the vignette:

> 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.

Subjects were asked to indicate how much praise Klaus deserved (on a scale from 0 to 6). They were then asked to indicate whether Klaus was *trying* to hit the device and whether Klaus *intentionally* hit the device.

On the second page, subjects received a 'skill' version of the same vignette. The skill version was exactly the same as the no-skill version except that the third paragraph was replaced with:

> With that thought in mind, he raises his rifle, gets the device in his sights, and presses the trigger. Klaus is an expert marksman. His hands are steady. The gun is aimed perfectly…

Subjects then answered the question: 'In this second version of the story, how much *praise* does Klaus deserve for what he did?' Here again, subjects marked the amount of praise on a scale from 0 to 6.

Overall, subjects gave the agent a lot of praise in the no-skill version (M = 4.46) and only slightly more in the skill version (M = 5.01), leaving a fairly low level of dissociation (M = .56). In other words: what we have here is a no-skill behavior such that people don't blame the agent but the degree of dissociation is fairly low. Thus, the dissociation hypothesis and the blame hypothesis generate competing predictions. The dissociation hypothesis predicts that people will

regard the behavior as intentional; the blame hypothesis predicts that people will regard the behavior as unintentional

In fact, most subjects (92%) said that the behavior was intentional. This result creates serious problems for the blame hypothesis. (How could it be that these people's judgements were influenced by blame, given that they specifically regarded the behavior as *praiseworthy*?) At the same time, it provides another piece of supporting evidence of the dissociation hypothesis.

To get a better sense for the overall pattern of results, we can combine the data from the present experiment with the data from experiment 2. For each subject, we can calculate a 'valence score' by treating ratings of praise as positive numbers and ratings of blame as negative numbers. Then we can calculate a 'dissociation score' by subtracting the rating of praise or blame in the skill version from the rating of praise or blame in the no-skill version. Finally, we can check to see how well each of these scores predict people's judgements as to whether or not the behavior is intentional. The results are displayed in the table below:

	% Intentional	Valence	Dissociation
Achievement	23%	3.4	1.7
Immoral	91%	-5.4	.3
Morally Good	92%	4.5	.6

Note, first of all, that the valence scores don't stand in any clear relation to people's intentional action judgements. People seem to regard the behavior as intentional not only when they give the agent blame (in the immoral condition) but also when they give the agent praise (in the morally good condition). And yet, in the achievement condition, where people also give the agent praise, they somehow conclude that the behavior is unintentional.

By contrast, when we turn to the dissociation scores, a clear pattern emerges. Both the immoral behavior and the morally good behavior yield low levels of dissociation, and both are regarded as intentional. The achievement behavior shows a high level of dissociation, and people regard it as unintentional. Collapsing across all three conditions, we find that dissociation is correlated with people's intentional action judgements, $r = -.39$, $p < .001$. The dissociation hypothesis thereby receives considerable support.

Conclusion

Assuming (at least for the sake of argument) that the dissociation hypothesis has been shown to be correct, let us now ask what significance this finding might have for the broader questions about people's concept of intentional action.

We began by describing two opposing viewpoints about the function that this concept serves in people's lives. The first viewpoint said that the concept should be understood as an element in a theory whose ultimate purpose is to predict and explain human behavior; the second viewpoint claimed that the concept was bound up in a fundamental way with *evaluative* tasks, such as the assignment of praise and blame. Hence, the first viewpoint says that the concept should only have features that make it well-suited to the tasks of prediction and explanation, whereas the second viewpoint says that the concept should also have features that make it well-

suited to the task of *evaluating* behavior — e.g., of assigning praise and blame — even when such features don't contribute at all to people's predictive or explanatory abilities.

I now want to argue that our findings count strongly against the first view and in favor of the second. That is to say: I want to argue that the features we have uncovered do not make the concept of intentional action well-suited to the tasks of prediction and explanation but do make the concept well-suited to the task of evaluating behavior.

Let us begin by considering a simple example. Suppose that we are observing a hunter who is trying to shoot a deer. The hunter is extremely unskilled, but — through sheer luck — he manages to hit the deer directly in the heart. Our task now is to determine whether or not the hunter killed the deer *intentionally*. According to the hypothesis defended here, our judgement will be influenced by the degree to which we believe that the hunter deserves reduced praise or blame because of his lack of skill. Thus, if some of us see the killing of deer as morally wrong and others see it as an achievement, the final result will probably be that some of us see the behavior as intentional and others do not.

It is hard to see any way in which this difference in our judgements could help us to pursue 'scientific' tasks like predicting future behavior. On the contrary, it seems that we should be able to reach perfect agreement about how our observation of the hunter should affect predictions of his future behavior even if we disagree radically about the moral status of the event we have observed.

If, however, one regards the concept of intentional action as a tool that is specifically well-suited to the task of assigning praise and blame, the difference in judgements begins to make sense. Some of us need to determine how much blame the hunter deserves for his 'immoral' behavior; others need to determine how much praise he deserves for his

‘achievement.’ Given this fundamental difference in the evaluative questions we face, it makes perfect sense that we end up making different use of our information about the hunter’s skill. After all, even if we all agreed about how this information should affect our predictions of future behavior, we would nonetheless have quite strong disagreements about how the information ought to affect the praise or blame that we assign to the agent.

Ultimately, then, the concept of intentional action may be best understood as something like a multi-purpose tool (Wilkes 1981). True, it can be used in the prediction of behavior, but it can also be used in evaluative tasks, such as the assignment of praise and blame. And all of these tasks — the evaluative no less than the predictive — play a role in shaping the concept itself.

References

ASTINGTON, J. W. (1999). The language of intention: Three ways of doing it. In P. D. ZELAZO, J. W. ASTINGTON, AND D. R. OLSON (Eds), *Developing theories of intention*. Mahwah, NJ: Erlbaum.

ASTINGTON, J. W. (2001). The paradox of intention: Assessing children's metarepresentational understanding. In B. F., MALLE, L. J. MOSES, & D. BALDWIN (Eds), *Intentions and intentionality: Foundations of social cognition*. Cambridge, MA: M. I. T. Press.

BRATMAN, M. E. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

CHURCHLAND, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*, 67-90.

CHURCHLAND, P. M. (1991). Folk psychology and the explanation of human behavior. In J. GREENWOOD (Ed.), *The future of folk psychology: Intentionality and cognitive science*. Cambridge: Cambridge University Press.

GOPNIK, A., & WELLMAN, H. M. (1992). Why the child's theory of mind really *is* a theory. *Mind and Language, 7*, 145-171.

GOPNIK, A., & MELTZOFF, A. (1997). *Words, Thoughts and Theories*. Cambridge, MA: MIT Press.

HARMAN, G. (1976). Practical reasoning. *Review of Metaphysics, 29*, 431-463.

LOWE, E. J. (1978). Neither intentional nor unintentional. *Analysis, 38*, 117-118.

MALLE, B. F. & KNOBE, J. (1997a). The folk concept of intentionality. *Journal of Experimental Social Psychology* 33: 101-121.

MALLE, B. F. & KNOBE, J. (1997b). Which behaviors do people explain? A basic actor-observer asymmetry. *Journal of Personality and Social Psychology, 72,* 288-304.

MALLE, B. F., KNOBE, J, O'LAUGHLIN, M. J., PEARCE, G. E., & NELSON, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions. *Journal of Personality and Social Psychology*, *79,* 309-326.

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. R. & MOSER, P. K. (1994). Intentional action. *Nous*, *28*, 39-68.

MELE, A. R. & SVERDLIK, S. (1996). Intention, intentional action, and moral responsibility. *Philosophical Studies*, *82*, 265-287.

PITCHER, G. (1970). 'In intending' and side effects. *Journal of Philosophy*, *67*, 659-668.

SHAVER, K. (1985). *The attribution of blame: Causality, responsibility and blameworthiness*. New York: Springer.

WILKES, K. (1981). Functionalism, psychology, and the philosophy of mind. *Philosophical Topics*, *12,* 1.

YOUNG, M. (2001). 'It's the thought that counts': The role of perceived intention in making event explanations. MS. Stanford University.