FOLK PSYCHOLOGY, FOLK MORALITY

Joshua Michael Knobe

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF PHILOSOPHY

November 2006
Abstract

The phrase ‘folk psychology’ refers to the ordinary psychology we use to understand each other and ourselves. Although folk psychology may initially seem to be a rather simple sort of thing, it can actually be surprisingly difficult to figure out precisely how it works.

One common view is that folk psychology is best understood as a tool for predicting and explaining behavior. The key idea here is that, if we attribute mental states to other people, we will be able to do a better job of predicting and explaining the behaviors they subsequently perform. Philosophers who hold this view often suggest that folk psychology is something like a scientific theory.

The present dissertation argues for a radically different view. It suggests that folk psychology might be better understood as a kind of multi-purpose tool. Specifically, the claim is that folk psychology should be understood as a tool not only for generating predictions and explanations but also for generating moral judgments.

The dissertation begins with a general theoretical defense of this approach. It then proceeds to apply the approach to three major aspects of folk psychology.

The first of these is people’s ordinary concept of intentional action. A series of experiments demonstrate that people’s attributions of intentional action in particular cases can actually be influenced by their moral beliefs. This result suggests that moral considerations may actually be playing a role in the concept of intentional action itself.

The second aspect is our ordinary practice of reason explanation. Here again, a series of experiments demonstrate the role of moral considerations and thereby suggest that morality may play a role in the relevant concepts.

The final aspect to be considered is people’s practice of causal attribution. Moral considerations have long been known to play a role in this practice; the claim defended here is that this role is best understood in terms of the nature of the concept of causation itself.
Table of Contents

Acknowledgements

A Brief Introduction

I. Folk Psychology: Science and Morals

II. Intentional Action

III. Reason Explanations

IV. Causal Attributions

References
Acknowledgements

A great many people offered me invaluable help with this dissertation, but I would especially like to thank John Doris, Gilbert Harman, Sean Kelly, Alan Leslie, Bertram Malle, Alfred Mele, Thomas Nadelhoffer, Shaun Nichols and Stephen Stich. I am grateful to them not only for their many helpful comments on the ideas discussed here but also — and this is surely far more important — for their pioneering work in developing the kind of philosophy practiced here.

I am also grateful in a different way to my wife Alina Simone and my friend Dan Moller for persuading me (against my initially fierce refusal) that I really ought to be writing on these topics.
A Brief Introduction

It is widely agreed that folk psychology plays an important role in people’s moral judgments. For a simple example, take the process by which we determine whether or not an agent is morally blameworthy. Although the judgment here is ultimately a moral one, it seems that one needs to use a fair amount of folk psychology along the way. Thus, one might determine that an agent broke the vase intentionally and therefore conclude that she is blameworthy for breaking it. Here it seems that one starts out with a folk-psychological judgment (that the agent acted intentionally) and then uses it as input to a process that eventually yields a moral judgment (that the agent is blameworthy). Many other cases have a similar structure.

In recent years, however, a number of studies have shown that there are also cases in which the arrow of causation goes in the opposite direction. That is, there appear to be cases in which people start out with a moral judgment and then use it as input to a process that eventually yields a folk-psychological judgment (Knobe 2003a, 2003b, 2004, 2005a, 2005b). These findings come as something of a surprise, and it can be difficult to know just what to make of them.

My own view is that the findings are best explained by the hypothesis that moral considerations truly do play a role in people’s underlying folk-psychological concepts (Knobe 2003b, 2004, forthcoming). The key claim here is that the effects revealed in recent experiments are not the result of any kind of ‘bias’ or ‘distortion.’ Rather, moral considerations truly do figure in a fundamental way in the issues people are trying to resolve when they grapple with folk-psychological questions.

I must confess, however, that not all researchers in the field share this view. Although many have been convinced that moral considerations actually do play a role in folk-psychological concepts, others have suggested that there might be better ways to account for the results of recent experiments. What we are left with, then, is an increasingly complex debate. Critics of my original proposal have constructed alternative hypotheses that seem to account for all of the data without assigning any fundamental role to moral considerations. Defenders then conduct new experiments that appear to falsify these alternative hypotheses. But the critics inevitably respond by constructing
even more sophisticated alternative hypotheses that manage to explain all of the new data while still assigning no fundamental role to moral considerations. And so the debate continues, with each new iteration yielding new levels of technical sophistication.¹

But lurking behind all the technical details, there seems to be a broader and more basic issue. The critics sometimes appear to feel that moral considerations just couldn’t be playing a fundamental role in folk psychology. The feeling is that, independent of the merits of any particular alternative explanation, one can tell that there must be some way to construct a valid alternative. This feeling is never articulated explicitly. Still, it comes through in the palpable sentiment that my defenders and I are upholding an absurd view and that we had really better come back to our senses.

In the first chapter, I try to confront that sentiment head on. My concern there is with the question as to whether we have any general theoretical reasons to expect that moral considerations will not play any fundamental role in folk-psychological concepts. Then, in the remaining chapters, I consider a number of different specific aspects of folk psychology and argue that moral judgments play a fundamental role in each of them.

Folk Psychology: Science and Morals

Much of the attractiveness of the view that moral judgments play no real role in folk psychology appears to stem from a perceived analogy between folk psychology and science. This idea is never spelled out explicitly, but the underlying argument seems to run something like this:

(1) Folk psychology is similar in many ways to a scientific theory.

(2) Scientific theories do not classify objects based on their moral properties.

We therefore have good reason to suppose that:

(3) Folk psychology does not classify objects based on their moral properties.

Of course, this is a not deductively valid argument, but it is a powerful one all the same. Both of the premises seem initially plausible, and together they appear to provide strong evidence for the conclusion.

To get a sense for the basic idea behind premise (2), it may be helpful to consider an example. Suppose we were able to observe a team of physicists studying the trajectories of certain projectiles. We might expect them to classify a projectile in terms of its mass, velocity, direction, and so forth. But suppose we then discover that their judgments can actually be influenced in some subtle way by moral properties, so that they sometimes end up applying scientific concepts to a projectile differently depending on whether they believe that it was morally right or morally wrong to launch it in the first place. In such a case, we surely would not conclude that moral properties actually play some important role in the basic concepts of physics. Instead, we would assume that the physicists were subject to some kind of bias that distorted their scientific judgment.

In thinking about cases like these, we brush up against some difficult questions about the relationship between science and morals. Someone might argue that initial impressions are deceiving here and that there really is some subtle sense in which scientific theories end up classifying objects on the basis of their moral properties. Perhaps there actually is something to this charge, but let us put it to the side for the moment. For the sake of argument, we can simply assume that scientific theories do not classify objects on the basis of their moral properties. Then we can go on to ask what
implications this putative fact about scientific theories might have for the study of folk psychology.

The key move, then, is from the claim that moral considerations are excluded from certain aspects of scientific theorizing to the claim that moral considerations are excluded from parallel aspects of folk psychology. This move rests on a certain analogy between science and folk psychology. The view is that, although science is more rigorous, more systematic and more explicit, we have reason to expect that the most basic practices associated with science will be found in folk psychology as well.2

It is this view that I want to examine here. To address these issues, we need to look more closely at the role science plays in people’s lives and the factors that have made it such a dominant approach to systematic inquiry. Then we can check to see whether those same factors can be found in the case of folk psychology or whether folk psychology differs from science in some important respect.

I

Contemporary enthusiasm for the analogy between folk psychology and science appears to stem, at least in part, from the extremely salient position that science occupies in modern life. Everywhere one looks, one finds the fruits of scientific inquiry, and it is easy to find oneself thinking that the practices we now associate with science are in some way ‘natural’ to human beings. One almost finds it difficult to imagine any other way of generating predictions or explanations.

But, of course, the matter is not so simple. Many of the practices that we now associate with science arose in a particular cultural context in the not-too-distant past. These practices are now quite widespread, but one cannot therefore infer that they reflect anything fundamental about human nature. It may well be that they only came to occupy

2 The idea that folk psychology resembles a scientific theory was perhaps first developed by Churchland (1981) and then came to play an important role in ‘theory of mind’ research as a result of work by Gopnik and Wellman (1992), Gopnik and Meltzoff (1997), and others. Since that time, numerous researchers have argued that folk psychology does not use quite the same methods that we find in scientific inquiry, but almost all of these researchers have assumed that folk psychology should be understood in terms of the typical goals of science (prediction, explanation, etc.). In a number of recent works, however, this assumption has been called into question. See especially Andrews (2006), Hutto (2004), Morton (2003) and Wilkes (1981).
such a salient position in our society because they do such a good job of solving the kinds of problems we most often encounter in modern life.

Perhaps some of the confusion here arises from our tendency to lump together a diverse array of practices and label them all collectively as ‘science.’ Some of the practices that fall under this label really do seem to reflect something fundamental about human nature. These practices can be found in young children and in people from other cultures, and many cognitive scientists believe that they have an innate basis (see, e.g., Bloom 2004; Gopnik et al. forthcoming; Keil 1989; Pinker 1997). But not all of the practices associated with science work like that. Some of them were only developed in recent centuries and appear to be passed down from one generation to the next through explicit instruction. There is little reason to suppose that these practices reflect anything fundamental about our innate cognitive endowments (Faucher et al. 2002; McCauley 2000).

The thing to keep in mind in discussing practices of this latter type is that they arose as a result of certain contingent historical events. There is an important sense in which the ‘scientific revolution’ of the 16th and 17th centuries truly was a revolution. It introduced genuinely new practices, practices that cannot be found in earlier eras. These practices subsequently assumed a dominant role in the kinds of inquiry conducted in systematic research programs, and we have ample evidence that they do a wonderful job of helping us get at the truth about certain difficult questions. But it would be wrong to suppose that there is something basic about human nature that compels us to adopt these practices in the form in which they presently exist. At other times and in other cultures, people have generated predictions using approaches that differed in various ways from the approach we now associate with science.

With this background in place, we can return to our central question. That question was whether we have any general theoretical reason to suppose that folk psychology treats moral considerations in the same way that science does.

II

The idea that folk psychology might be similar to science has been encouraged by the claim that folk psychology should be understood (in a certain technical sense) as a
theory. The association here is understandable. As soon as one hears the word ‘theory,’ one immediately thinks of the sciences. So when one is told that folk psychology itself should be understood as a theory, one naturally leaps to the conclusion that folk psychology should be understood as something like science. It is therefore essential to remember that the word ‘theory’ was first introduced into this discussion in a highly specialized sense that did not carry any implications about all of the practices we normally associate with science.

The idea that folk psychology should be understood as a theory was first developed by Sellars (1956) and then entered the world of cognitive science through the influential work of Premack and Woodruff (1978). These researchers were concerned with the fact that folk psychology doesn’t just give us a collection of empirical generalizations about observable phenomena but actually provides a deeper sort of account that works by explaining observable behaviors in terms of unobservable mental states. As Premack and Woodruff put it:

In saying that an individual has a theory of mind, we mean that the individual imputes mental states to himself and to others… A system of inferences of this kind is properly viewed as a theory, first, because such states are not directly observable, and second, because the system can be used to make predictions, specifically about the behavior of other organisms. (Premack & Woodruff 1978:515)

I have no objections to this use of the term ‘theory,’ but when the term is used in this way, one cannot simply assume that every theory is best understood on the model of science. After all, a system of belief can easily qualify as a ‘theory’ in Premack and Woodruff’s sense even if it does not have many of the properties we normally associate with scientific inquiry. To take a particularly glaring example, certain religions posit unobservable entities that can be used to predict observable events and might therefore be described as ‘theories.’ Now, it does seem fair to say that a religion can offer us a theory about how the world works, but one sees immediately that the theories offered by religions differ from scientific theories in a number of important respects.

In particular, the argument sketched above seems to depend in a crucial way on the distinctive features of scientific theories. There is some intuitive plausibility to the inference: ‘Folk psychology is similar to science. Therefore, it does not classify objects
based on their moral properties.’ But the argument loses all its force when we change it to: ‘Folk psychology is similar to religion. Therefore, it does not classify objects based on their moral properties.’ Religions serve a great many different functions in our lives, and prediction is just one. No one would be surprised to find that religious theories are connected in an essential way with moral considerations.

In short, it is easy to get confused by the claim that folk psychology is a ‘scientific theory.’ We really need to divide this claim into two parts — the claim that folk psychology is a theory and the claim that folk psychology is scientific. The claim that folk psychology is a theory simply isn’t very relevant to the questions we are trying to address here. What we really want to know is whether folk psychology is, in the relevant sense, scientific.

III

Our concern, then, is with the distinctive features of scientific theories — the features that distinguish scientific theories from theories of other types. It seems that these features lie not so much at the level of content as at the level of methodology. The methods we use to evaluate scientific theories seem to differ in some important respects from the methods we use to evaluate theories of other types.

Perhaps the most striking aspect of scientific methodology is its sensitivity to empirical evidence. We use scientific theories to generate predictions which can then be tested through observation or experiment. Theories that yield false predictions may be revised or abandoned. So one way to determine whether folk psychology is something like a scientific theory would be to ask whether it, too, is sensitive in the right way to empirical evidence.

A whole industry of research has arisen to answer this question, and a wide variety of competing theoretical frameworks have now been proposed. Some have argued that people can revise the basic framework of folk psychology using the very same psychological processes that scientists use to revise their theories (e.g., Gopnik & Wellman 1992); others argue that the basic framework underlying folk psychology is innate and is only sensitive to empirical considerations through a process of evolution by natural selection (e.g., Baron-Cohen 1995); and still others have suggested that folk
psychology might be subserved by an innate module that uses empirical evidence to set
certain highly specific parameters (Stich & Nichols 1998). The debate among these
various positions is still ongoing.

But I worry that this research does not really get at the question we are trying to
address here. It is not as though scientific theories are the only systems of thought that
prove sensitive to empirical considerations. One finds at least some level of sensitivity to
empirical considerations even in systems of thought that are clearly non-scientific.
Consider a simple example. In the seventeenth century, many European Jews believed
that Shabbatai Zvi was the messiah. They then received a shocking piece of
disconfirming empirical evidence (Shabbatai Zvi converted to Islam), and most of them
soon abandoned their previous belief. What we have here is a clear case of a group of
people revising their views in light of empirical evidence. But no one would suggest that
the followers of Shabbatai Zvi were propounding a genuine scientific hypothesis!
Clearly, their belief was a religious doctrine, and the criteria used to evaluate it therefore
differed quite radically from the criteria we typically find in scientific inquiries.

The key mistake here is to assume that we can figure out what is special about
scientific inquiry simply by looking at the considerations that scientists normally take
into account. This approach has undoubtedly yielded many important insights, but it is
not sufficient all by itself. We also need to look for kinds of considerations that scientists
don’t take into account. That is, we need to look for kinds of considerations that figure
prominently in other systems of thought but do not play any role in scientific inquiries.

To get a sense for what I mean here, consider the many kinds of criteria we might
use in deciding between competing religious doctrines. It seems that many of these
criteria play no role at all in scientific investigations. Indeed, one of the key turning
points in the scientific revolution was the struggle to establish a special realm of inquiry
from which these other criteria would be completely excluded.

For present purposes, one of the most important distinctions between scientific
and non-scientific theories lies in the differing roles they assign to moral considerations.
We expect a religious doctrine to give us some measure of moral guidance, and if it fails
to do so, we regard it as deficient in an important respect. By contrast, when we are
evaluating a scientific theory, it seems that we are not supposed to be concerned in an
essential way with moral questions. The theory can be perfectly successful from a scientific point of view even if it provides no moral guidance at all. In fact, we might find that the theory carves up the phenomena in a way that is completely orthogonal to the categories that prove most relevant in our moral thinking. But no matter. As long as the theory does well according to the distinctive criteria of science (empirical adequacy, simplicity, etc.), we are supposed to consider it a success.

We can now get a better handle on the question as to whether or not folk psychology is something like a scientific theory. In addressing this question, it is not enough just to ask whether or not folk psychology is sensitive in the right way to the kinds of considerations that play a role in scientific inquiry. We also need to know whether it resembles science in excluding the kinds of considerations that are usually excluded from scientific inquiries.

IV

At this point, it might be thought that we really do have quite good reason to assume that folk psychology excludes the very same sorts of considerations that are normally excluded from scientific inquiries. After all, it is a conspicuous fact about our modern age that scientific approaches have proved extraordinarily successful in the systematic research programs where they are most commonly employed. One might therefore be tempted to conclude that the most effective way to proceed as folk psychologists would be to use almost exactly the same methods we find used in scientific inquiry.

But perhaps this conclusion is a bit too hasty. Clearly, there are some important differences between what we are looking for in a scientific research program and what we are looking for in a folk theory like folk psychology. So it is at least conceivable that the approach that best serves our needs in scientific research programs will not also best serve our needs in folk theories. Before we can determine whether or not there is reason to suspect that folk psychology uses a scientific approach, we therefore need to look in more detail at the advantages and disadvantages of that approach more generally.

One of the chief advantages of the scientific approach is its unparalleled predictive power. By excluding many of the criteria used in other kinds of inquiry, a
scientific investigation can arrive at theories that do an extraordinarily good job at predicting the phenomena under study.

But this predictive power comes with a price. A scientific theory is a highly special-purpose tool. It might do an excellent job when our aim is to make predictions, but it won’t necessarily prove helpful in all of the other tasks for which we ordinarily use complex conceptual thought. In particular, it won’t necessarily carve up the phenomena in a way that proves helpful for making moral judgments.

Think, for example, of the various ways in which we might divide people up into categories. One approach would be to develop concepts that did the best possible job of predicting and explaining behavior. (And here we might end up with concepts like *person with high serotonin levels.*) But the categories we construct using this approach may turn out not to be ideal when it comes time to make moral judgments. Indeed, it may turn out that the categories that prove most helpful in making moral judgments are completely orthogonal to the categories that prove most helpful in generating predictions and explanations.

Assuming that we do want to make moral judgments, it seems that we will need to develop additional, non-scientific concepts that help us to pick out the morally relevant categories. Ultimately, we will then be left with two different ways of carving up the same class of phenomena. We will have concepts that pick out the categories that prove most helpful in prediction and explanation (e.g., person with high serotonin levels) and also concepts that pick out the categories that prove most helpful in making moral judgments (e.g., morally good person). We will then need a complex system of rules that enables us to move from one set of concepts to the other.

For cognitively limited creatures like ourselves, this level of specialization might be a major problem. We would have to retain in our minds two distinct systems of concepts, two distinct kinds of psychological mechanisms, two distinct sets of propositional attitudes. Whenever we were engaged in tasks that involved both prediction and moral judgment, we would have to shift back and forth from one system of categories to the other. All this would impose a substantial demand on our cognitive resources.
In short, the sort of approach we now associate with science has both advantages and disadvantages. The chief advantage lies in its *predictive power*; the chief disadvantage lies in the resulting *conceptual complexity*.

V

There is, however, another possible approach. Instead of having one system of concepts for use in generating predictions and then a second, completely separate system of concepts for use in making moral judgments, we could have a single system of concepts that was used for both of these tasks. This single system of concepts might not do a perfect job either at generating predictions or at making moral judgments, but it could do at least an adequate job of both. Hence, although this system of concepts might not afford us the greatest possible predictive power, it would do quite a bit to reduce the amount of cognitive complexity we needed to handle.

For an analogous case, consider the various ways we might come to conceptualize the weather. In thinking about the weather, there is a need to make *predictions* about what conditions will arise in the future, and there is also a need to make *evaluations* of whether these conditions are good or bad for certain purposes. What sorts of concepts would best enable us to achieve these goals? One approach would be to have a system of concepts that was specifically suited to the task of making predictions and then another, entirely separate system of concepts that was specifically suited to the task of evaluation. But such an approach might leave us with a large and unwieldy array of distinct ways of carving up the same class of phenomena. We might therefore be better served by a single system of concepts that wasn’t ideally suited either for prediction or for evaluation but could serve us at least fairly well in both of these tasks.

It is certainly conceivable that folk psychology uses a system of concepts that works more or less along these lines. That is, it is conceivable that folk-psychological concepts are constructed in such a way that they do an adequate job at helping us both with prediction and with moral judgment, though perhaps without doing an absolutely ideal job in either of these two domains. What we want to know now is whether there are any general theoretical arguments against the view that folk psychology works in this way.
Thus far, we have been considering one possible argument. This argument relies on an analogy between folk psychology and systematic science. It points out that systematic scientific research programs typically don’t try to develop a small set of concepts that enable us to do at least passably well at a wide variety of different tasks. Instead, they typically seek to develop concepts that enable us to do the best possible job at a specific range of tasks (prediction, explanation, etc.), even if they thereby end up coming up with concepts that aren’t especially helpful in the task of making moral judgments. The argument then suggests that this fact about the concepts used in systematic science gives us reason to expect to find something similar in the concepts used in folk psychology.

At least for the sake of argument, we have been accepting all of the relevant claims about the nature of systematic science. The key question then becomes whether these claims can justify the relevant inferences about folk psychology.

VI

But when the question is put in these terms, one notes immediately that folk theories are quite different from the sorts of theories one typically develops in systematic scientific research programs. Clearly, the two kinds of theories occupy two very different kinds of roles in our lives, and there is therefore little reason to expect that people look to them to fulfill the same needs. Most importantly for present purposes, it seems that people are far more reluctant to tolerate conceptual complexity in a folk theory than they are in the theories they employ in systematic research programs.

In systematic research programs, one can easily deal with the problem of conceptual complexity through a division of cognitive labor. No individual researcher needs to learn all of the scientific concepts; each only needs to know the concepts used in one particular domain of inquiry. Thus, science as a whole can acquire an extraordinary level of conceptual complexity even without any individual person grasping more than a tiny fraction of the total.

This solution is not available in the case of folk psychology. We cannot make do with a system in which one person only knows the emotion concepts, another only knows the trait concepts, and so on. We will only be able to do a tolerable job of getting around
in the world if each person has some grasp of the whole of folk psychology. In fact, this seems to be one of the fundamental differences between folk theories and systematic research projects. We do not look to folk theories for a system that can serve, at least in principle, to generate perfectly accurate predictions. We look to them for tools that can help creatures like us — with all of our cognitive limitations — to accomplish certain practical goals.

Ultimately, then, it seems that we have good reason to expect that the concepts used in folk psychology will differ in certain respects from the concepts used in systematic research. In systematic research projects, one should expect to find an enormous array of different concepts, with each concept highly specialized for one particular use. But there is good reason to expect that folk theories will work somewhat differently. In a folk theory, one should expect to find concepts that are less highly specialized and can therefore be used in a wider variety of different tasks. Each concept might be specific to one particular domain of phenomena, but it will be constructed in such a way as to help us do almost anything we might want to do with the phenomena in that domain. Thus, instead of expecting to find a clear distinction between concepts used for prediction and concepts used for moral judgment, one should expect to find concepts that are not specialized for either of these two tasks but are constructed in such a way that they can do a decent job of both.

VII

There seems to be a widely shared intuition that moral considerations just couldn’t be playing any fundamental role in the basic concepts of folk psychology. Researchers who hold this intuition have not backed it up with systematic arguments. In fact, they have not even discussed it explicitly. Yet the underlying intuition comes through quite clearly in the incredulous stares one receives whenever one suggests that some particular folk-psychological concept might be best understood as having moral features.

My concern here has been with the question as to whether there actually are any general theoretical arguments in favor of this intuitive view. I focused in particular on the argument that we have reason to expect that folk psychology will show certain
fundamental similarities to scientific inquiry. This argument did not fare especially well on closer inspection. In fact, it seems that we actually have some reason to expect that folk psychology will differ from science in the relevant respects.

Of course, it is possible that there really are good arguments for the view that moral considerations can’t play any fundamental role in folk-psychological concepts and that these arguments have simply eluded my grasp thus far. In that case, I would want to know exactly what the relevant arguments are. Clearly, we should not reject a hypothesis simply because it goes against our philosophical preconceptions. What we need now are definite theoretical proposals that generate testable predictions about the structure of people’s folk-psychological concepts.
Intentional Action

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 idea 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).

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.

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

Let us begin by considering one simple case:

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

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

They started the new program. Sure enough, the environment was harmed.
Now ask yourself: Did the chairman of the board intentionally harm the environment?

Faced with this question, most people (though certainly not all) say that the
answer is yes. And when asked why they think that the chairman intentionally harmed the
environment, they tend to mention something about the chairman’s psychological state —
e.g., that he decided to implement the program even though he specifically knew that he
would thereby be harming the environment.
But it seems clear that these facts about the agent’s psychological state cannot be all there is to the story. For suppose that we replace the word ‘harm’ with ‘help,’ so that the vignette becomes:

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

The chairman of the board answered, ‘I don’t care at all about helping the environment. I just want to make as much profit as I can. Let’s start the new program.’

They started the new program. Sure enough, the environment was helped.

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

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

Of course, it would be a mistake to base such a broad claim on evidence from just one vignette. But the claim becomes plausible when one sees how robust the effect is. The effect continues to emerge when the details of the story are switched around (Knobe 2003a); it emerges when intuitions are measured using a quite different methodology (Knobe 2004); it emerges when the whole experiment is translated into Hindi and run with Indian subjects (Knobe & Burra forthcoming); it even emerges when subjects are only 4 years old (Leslie & Knobe 2004). Moreover, philosophers have constructed other,
very different cases in which moral considerations appear to influence people’s intuitions about whether or not a given behavior is intentional (Harman 1976; Lowe 1978), and when these other kinds of cases have been put to an experimental test, the effect emerges on them as well (Knobe 2003b; Nadelhoffer forthcoming a).

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

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

II

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

In describing the factors that influence people’s intuitions, it will often prove helpful to make reference to the various features that philosophers have discussed in their analyses of the concept of intentional action. Here we shall be principally concerned with the features trying, foresight and skill. There has been a great deal of controversy in the philosophical literature about the role that each of these features plays in the concept of intentional action (for an excellent review, see Mele 1992). In the present context,
however, it will not be necessary to discuss these controversies in any real detail. Instead, what we want to show is that, in the cases under dispute, peoples intuitions are influenced by the moral status of the behavior.

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

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

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

This is the key result of the experiment described above — where a vignette about environmental harm elicited very different intuitions from a quite similar vignette about environmental help. And the same effect arises for other cases that have the same basic structure. So, for example, when we transpose the story from a corporate boardroom to a battlefield — with a lieutenant helping or harming his troops in place of a chairman helping or harming the environment — we still get the same basic effect. People say that the lieutenant acted intentionally if he harmed the troops as a side-effect but that he did not act intentionally if he helped the troops as a side-effect (Knobe 2003a).
Cases of side-effects are not the only ones in which moral considerations play a role. Similar issues arise in cases where the agent lacks skill. Consider a case in which an agent is trying to perform a behavior and actually does succeed in performing that behavior. And now suppose that the agent didn’t really have the skill to perform that behavior in any reliable fashion, so that ultimately the agent only manages to succeed through sheer luck. Has the agent performed the behavior intentionally? According to some philosophical analyses, the answer is yes (e.g., Brand 1984); according to others, the answer is no (e.g., Mele & Moser 1994). But once again, it appears that neither view correctly predicts people’s intuitions in all cases. People’s intuitions about these cases seem to depend in part on the moral status of the behavior itself.

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

Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bulls-eye. He raises the rifle, gets the bull’s-eye in the sights, and presses the trigger.

But Jake isn’t very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…

Nonetheless, the bullet lands directly on the bull’s-eye. Jake wins the contest.

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

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

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

But Jake isn’t very good at using his rifle. His hand slips on the barrel
of the gun, and the shot goes wild…

Nonetheless, the bullet hits her directly in the heart. She dies instantly.

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

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

Klaus is a soldier in the German army during World War II. His regiment has been sent on a mission that he believes to be deeply immoral. He knows that many innocent people will die unless he can somehow stop the mission before it is completed. One day, it occurs to him that the best way to sabotage the mission would be to shoot a bullet into his own regiment’s communication device.

He knows that, if he gets caught shooting the device, he may be imprisoned, tortured or even killed. He could try to pretend that he was simply making a mistake — that he just got confused and thought the device belonged to the enemy — but he is almost certain that no one will believe him.

With that thought in mind, he raises his rifle, gets the device in his sights, and presses the trigger. But Klaus isn’t very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild…

Nonetheless, the bullet lands directly in the communications device. The mission is foiled, and many innocent lives are saved. Here most people feel that Klaus did hit the communications device intentionally.

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

Thus far, we have reported results from only two experiments. But these results have been replicated and extended in recent work by the philosophers McCann (2004), Nadelhoffer (forthcoming a; forthcoming b) and Sverdlik (2004), by the psychologists Malle (forthcoming) and Leslie (Leslie & Knobe 2004), and by the anthropologist Sousa (forthcoming). At this point, there can be little doubt that moral considerations have an impact on people’s use of the word ‘intentionally.’ The key remaining questions are about how this effect is to be understood.

III

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

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

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

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

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

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

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

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

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

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

The claim here is an important and initially plausible one, but recent research appears not to lend it much support. Nichols and Ulatowski (2006) gave subjects the vignette about the executive who harms the environment and then asked them to choose between the following options:

(i) “The chairman intentionally harmed the environment, and he is responsible for it.”
(ii) “The chairman didn’t intentionally harm the environment, but he is responsible for it.”

These options give subjects an opportunity to say that the act was unintentional while entirely avoiding unwanted implicatures. If a subject chooses (ii), she says that the behavior was unintentional but then immediately cancels the implicature about responsibility by explicitly asserting that the agent actually was responsible for what he did.
Yet subjects elected not to select this option. In fact, they were no more likely to say that the behavior was unintentional when given the opportunity to cancel the implicature than they were when simply asked to give a yes/no answer about whether or not the behavior was performed intentionally. This result suggests that the original findings were not, in fact, due to the influence of conversational pragmatics.

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

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

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

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

The commonsense view works something like this:

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

One problem with this commonsense view is that it offers no explanation for the fact that people’s moral judgments sometimes influence their intuitions as to whether or not a behavior was performed intentionally. Malle therefore proposes that the process sometimes works more like this:

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

If the process really does work like this, it would be reasonable to infer that people were making some kind of error. This model does not posit a role for moral considerations in the fundamental competence underlying people’s concept of intentional action. Rather it seems to be describing a kind of bias that can infect people’s thought processes and lead them astray.
There is, however, another plausible way to make sense of the data reported thus far. Perhaps the process actually works like this:

![Diagram showing relationships between bad, intentional, and blameworthy]

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

In the cases we have been discussing thus far, these competing models make identical predictions. Take the case of the corporate executive who harms the environment. Here we find that people both (a) classify the agent’s behavior as bad and (b) blame the agent for that behavior. Since people judge the case to be both bad and blameworthy, there is no obvious way to figure out which of these two judgments is influencing their intuitions.

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

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

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

And yet, people generally say that the executive intentionally decreased sales in New Jersey (Knobe & Mendlow forthcoming). This result spells trouble for any theory that tries to account for the role of moral considerations in terms of blame alone. What we have here is a case in which the agent is not considered blameworthy but in which people’s beliefs about good and bad are nonetheless influencing their intentional action intuitions. This kind of result cannot plausibly be explained in terms of people’s efforts to justify a prior judgment of blame. (After all, there is no blame here to justify!) The most plausible hypothesis seems to be that people’s judgments of good and bad are actually playing a role in the fundamental competencies underlying their concept of intentional action.

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

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

The answer lies, I think, in a particular view about the nature of folk psychology. This view says that the basic purpose of folk psychology is to enable people to predict each other’s behavior or to offer them some other form of quasi-scientific, purely naturalistic understanding. When folk psychology is understood in this way, it seems that it would be pointless for moral considerations to play any real role. Thus, if moral considerations appear to be influencing people’s use of words like ‘intentional,’ one is naturally led to search for some alternative to the view that these considerations are actually having an impact in the fundamental competencies underlying folk psychology. The goal then becomes to find some way in which people’s fundamental competencies can be overridden, corrupted or otherwise shielded from view.

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

IV

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

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

First of all, we should note that the three features we encountered in our discussion of intentional action — trying, foresight and skill — play a crucial role in the process by which people normally assign praise and blame. Thus, when people are wondering how much praise or blame an agent deserves, their conclusion will sometimes depend on whether or not the person was trying to perform a given behavior, whether she chose to do something that she foresaw would involve performing that behavior, whether she had the skill to perform that behavior reliably.

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

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

A behavior is shmintentional if and only if the agent had skill and either trying or foresight.
We could not make praise and blame judgments by simply checking to see whether a given behavior was shmintentional. The problem is that different features are relevant to different behaviors and that shmintentionality is therefore more relevant to praise and blame judgments for some behaviors than for others.

For a simple example, we can return to the environmental cases that we presented above. Suppose that an agent decides to perform a given behavior because he wants to increase profits. The agent knows that his behavior will have some impact on the environment. But he does not care at all about the impact he is having on the environment — he is only performing the behavior as a way of increasing profits. Will people feel that this agent deserves any praise or blame for what he has done? Clearly, people’s views will depend on the particular type of impact that the agent is having on the environment. If the agent is harming the environment, they may feel that he deserves a considerable amount of blame. But if he is helping the environment, they will probably feel that he deserves almost no praise.

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

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

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

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

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

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

\[
\begin{array}{ccc}
\text{bad?} & \text{intentional?} & \text{blameworthy?} \\
& \text{choose features} & \text{does behavior have features} \\
\end{array}
\]

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

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

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

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

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

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

People offer different types of explanations for different types of phenomena. The particular type of explanation we will be concerned with here can be illustrated with the following example.

(1) George decided to buy the Toyota because it got such good gas mileage.

This is an explanation of a quite distinctive sort. It is profoundly different from the sort of explanation we might use to explain, say, the movements of a bouncing ball or the gradual rise of the tide on a beach. Unlike these other types of explanations, it explains an agent’s behavior by describing the agent’s own reasons for performing that behavior. Explanations that work in this way have a number of distinctive and important properties, and we will refer to them here as reason explanations.

Looking at the use of reason explanations with a philosophical eye, one is apt to experience a certain puzzlement. One wants to know precisely what makes a given reason explanation true or false. So, for example, explanation (1) seems to be saying that George’s reason for deciding to buy the Toyota was that it got such good gas mileage. But what exactly makes it the case that this is George’s reason? Does he have to actually be thinking about it at the time he makes the decision? Does his thought have to be the cause of his decision? Does he have to know why he decides as he does? These questions lie at the heart of an ongoing philosophical debate about the nature of reason explanations.3

Thus far, research on these issues has made use of a somewhat peculiar method. Philosophers do not simply look at people’s intuitions and check to see which factors make people more or less likely to regard a given reason explanation as acceptable. Instead, the goal is to come up with a unified philosophical account of the nature of reason explanations themselves. Such an account would tell us, on a truly fundamental level, what reasons are and what role reason explanation plays in people’s lives. The assumption seems to be that, once we have the answers to these fundamental questions in

3 The philosophical literature on these issues is enormous. For seminal early discussions, see Wittgenstein (1958), Anscombe (1957) and Davidson (1963). For a few contemporary discussions, see Bittner (2001), Dancy (2000), Mele (forthcoming) and Schueler (2003).
hand, we will be able to derive from the underlying philosophical theory a series of conclusions about the exact criteria for determining whether or not a given reason explanation is acceptable.

This research program seems to me to be a valuable and important one, but I am not sure how much it can tell us about the actual practice of reason explanation. After all, it is not as though this practice was carefully engineered to do a perfect job of accomplishing some particular goal. Presumably, the practice arose through a more haphazard process, picking up a whole variety of arbitrary quirks along the way. Thus, if we want to understand the practice of reason explanation — the actual practice, the one we engage in every day — it is no use looking for a single unified philosophical account that can capture it all. We will have to come to grips with a variety of little rules that cannot be derived from any more general theory but just happened to arise at one time or another.

My concern here is with one of these little rules. Earlier studies have shown that people’s intuitions about whether a given behavior was performed for a reason can be sensitive to the moral status of the behavior itself. In other words, if we are wondering whether George decided to buy a Toyota for a particular reason, we will need to know whether the behavior of buying a Toyota was morally good or morally bad. My goal here is to understand how exactly this effect works and why it arose in the first place. Ultimately, the conclusion will be that the effect cannot be derived from any broader theory about the underlying function of reason explanations but that it is possible to make a certain kind of sense of it all the same.

I

The best way to convey the distinction between reason explanations and explanations of other types is simply to provide a few examples. Thus, consider the two explanations:

(2) Susan yelled at her son because he broke the vase.

(3) Susan yelled at her son because she is an irritable person.
Intuitively, it seems that there is a fundamental difference between these two explanations. One way to get a rough sense for the nature of this difference is to imagine a kind of inner monologue that the agent goes through before making a decision. The agent might think: ‘He broke the vase; therefore I will yell at him.’ But she definitely would not think: ‘I am an irritable person; therefore I will yell at him.’ Presumably, the thought that she is an irritable person does not enter into her practical reasoning in any way.

But, clearly, the agent does not literally go through an inner monologue of the type we have been describing. So perhaps it would be helpful to seek some other way of explaining what it is to offer a reason explanation.

One of the most appealing approaches here is to make use of a metaphor. In particular, one wants to describe the phenomenon using either the metaphor of a ground or the metaphor of a light. Thus, one might say that Susan yells at her son on the grounds that he broke the vase (whereas she does not yell at him on the grounds that she is an irritable person). Or, alternatively, one might say that she yells at her son in light of the fact that he broke the vase (whereas she does not yell at him in light of the fact that she is an irritable person).

In general, I find that most people can understand the distinction between reason explanations and explanations of other types after considering a few examples, imagining certain inner monologues, and mulling over the metaphors of ground and light. At the very least, people who are given these rough specifications have a pretty good sense for what sort of distinction philosophers were trying to make when they introduced the notion of ‘reason explanations.’ But this vague sort of understanding is not what one usually seeks when doing philosophy. One wants a more detailed specification of the exact conditions an explanation would have to meet before it counts as a reason explanation.

At this point, it might be thought that what we need is a stipulative definition. If we want to use the phrase ‘reason explanations,’ we can start out by simply specifying the precise conditions that explanation has to meet before this phrase can apply to it.
Then we can check to see what factors influence people’s use of the explanations that meet the stipulated conditions.

This does not seem to me to be the best way to proceed. There is, I think, no need for philosophers to construct their own stipulated definitions before looking at the kinds of explanations people ordinarily provide. Rather, it seems that people themselves distinguish between reason explanations and explanations of other types. The very rough characterization we have given in this section should be sufficient to pick out the distinction in question. What we want to know now is just whether moral considerations have any impact on the acceptability of the explanations that people themselves are classifying as reason explanations.

II

Of course, it may initially seem a bit absurd to suggest that people ordinarily distinguish between reason explanations and explanations of other types. After all, the phrase ‘reason explanation’ was first developed by philosophers, and people don’t often use it in everyday conversations. Nor do people seem to be in any way aware that they are classifying certain explanations as reason explanations. So when we say that people are distinguishing between different types of explanations, we certainly don’t mean that they are consciously engaged in a process of distinguishing or classifying. The claim is, rather, that people are classifying certain explanations as reason explanations in an entirely non-conscious way.

To get a sense for what we are claiming here, it might be helpful to consider the kinds of theories one typically finds in linguistics. Linguists often draw distinctions that people have no awareness of making, but there does seem to be considerable evidence that people truly are making these distinctions at a non-conscious level. For a simple example, consider the distinction between qualitative adjectives and relational adjectives. Clearly, the phrases ‘qualitative adjective’ and ‘relational adjective’ were first developed by linguists, and most ordinary people are in no way aware of making a distinction between the two categories. Still, there does seem to be evidence that people make the distinction. Thus, in (4)-(7), it seems that people are classifying ‘professorial’ and ‘paternal’ as qualitative adjectives and ‘financial’ and ‘regional’ as relational adjectives.
We can see that people are distinguishing between the two types of adjectives when we try to include each of them in a different role within the sentence. For example:

(8) The gentleman is professorial.
(9) The old man is paternal.
(10) *The management is financial.
(11) *The studies are regional.

The fact that people regard (8) and (9) as grammatical and (10) and (11) as ungrammatical gives us some evidence that people are distinguishing between the different types of adjectives. Though they are not consciously making the distinction, we have indirect evidence that they are making the distinction all the same.

In much the same way, we can have evidence that people are distinguishing reason explanations from explanations of other types. Here again, the evidence is indirect. It doesn’t come from asking people directly whether or not they distinguish different types of explanations. Rather, the claim is that we can explain a variety of observable phenomena using a theory according to which people non-consciously distinguish reason explanations from explanations of other types. These phenomena include the difference between the way people explain their own behavior and the way they explain the behavior of others, the techniques people use when providing explanations designed to make a good impression, the differences between explanations used for individuals and those used for groups (Knobe & Malle 2002; Malle, Knobe, O’Laughlin, Pearce & Nelson 2000; O’Laughlin & Malle 2002).

III

But now we face a problem. If we want to understand the conditions under which people regard reason explanations as acceptable, we need to have some way of determining whether or not people are classifying a given explanation as a reason explanation. Clearly, the difficulty we face here is far greater than the one we faced in the
previous chapter where we were concerned with the concept of intentional action. Our strategy for figuring out how people applied the concept of intentional action in a given case was simply to ask them. So, for example, when we wanted to know whether people classified the behavior of hitting the bull’s-eye as an intentional action, we simply asked them, ‘Did he hit the bull’s-eye intentionally?’ But this strategy is not available to us when we are trying to study the practice of reason explanation. We cannot simply present people with an explanation and ask them, ‘Is this a reason explanation?’ (Most people would have no idea what such a question even meant.) And so, we are forced to look for a more subtle method.

The obvious approach here would be to look at the linguistic formulations people use when offering reason explanations. Perhaps we can find particular phrases or particular syntactic structures that people use exclusively for reason explanations and not for explanations of other types. Then, if we find that a person is willing to express an explanation using one of these phrases or syntactic structures, we can be fairly sure that he or she has classified the explanation as a reason explanation.

This approach is a promising one, but one should note at the outset that it faces certain important difficulties. The problem is that people sometimes use exactly the same linguistic formulations even when they are giving explanations of very different types. Thus, consider the explanations:

(12) Because he so desperately wanted to win, he practiced long and hard.

(13) Because he so desperately wanted to win, his heart started beating rapidly.

These two explanations consist of exactly the same words ('because he so desperately wanted…'), and yet they are used in quite different ways. Explanation (4) gives the agent's reason for practicing long and hard, but explanation (5) does not give the heart's reason for beating rapidly — it simply cites a factor that caused the heart to beat as it did.

Still, there do seem to be linguistic formulations that are not ambiguous in this way. That is to say, there seem to be linguistic formulations that can only be used to give reason explanations and cannot be used in explanations of other types. Perhaps the clearest cases here are the purposive clauses — clauses of the sort that one might
introduce with expressions like 'for the sake of,' 'so that,' or 'in order to.' For a simple example, consider:

(14) He practiced long and hard in order to increase his chances of winning.

This linguistic formulation can only be used to give the agent's reason for practicing. It cannot be used to mean that the agent's desire to increase his chances of winning simply caused him to practice (without serving in any way as his reason for practicing). Thus, if we want to reach a better understanding of the conditions under which people are willing to use reason explanations, one helpful strategy would be to start out by looking at the conditions under which people are willing to use linguistic formulations like this one.

IV

With this basic framework in place, we can now turn to people’s intuitions about particular cases. What we wish to show is that these intuitions can sometimes be influenced by people’s beliefs about the moral status of the behavior in question.

We begin with a political example. One of Bill Clinton’s advisors comes to him and says: ‘We’ve just come up with a new campaign strategy. It will definitely increase your political popularity, but if you implement it, you’ll also be selling out the Left.’ Clinton answers: ‘I don’t care at all about what happens to the Left. All I want to do is increase my political popularity. Let’s use that new strategy!’ Now consider the explanation:

(15) Clinton sold out the Left in order to increase his political popularity.

Most people feel that this explanation sounds right. What I want to argue is that people only think it sounds right because they regard selling out the Left as a morally dubious behavior.

For some prima facie evidence in favor of this hypothesis, consider what happens when we replace selling out the Left with a behavior that has a different moral status. For example, suppose we change it to pleasing the people of Sweden. In this revised version, the advisor tells Clinton that the new strategy will increase his popularity and also please the people of Sweden. Clinton responds that he doesn’t care at all about the Swedes; all
he wants to do is increase his political popularity. Even if Clinton does thereby end up pleasing the people of Sweden, it sounds very wrong to say:

(16) Clinton pleased the people of Sweden in order to increase his political popularity.

My suggestion is that the key difference between explanations (15) and (16) lies in the differing moral status of the behaviors they explain.

Of course, there are also many other differences between selling out the Left and pleasing the people of Sweden. (One might argue, e.g., that selling out the Left is more important or that it is more relevant to the decision at hand.) To address these worries, I therefore ran an experiment using more closely matched cases. In fact, I used the very same pair of cases we encountered in the previous chapter — the pair of cases in which the chairman either helps or harms the environment. But even with this closely matched pair of cases, people’s intuitions showed a marked asymmetry. Most people thought it sounded right to say

(17) The chairman harmed the environment in order to increase profits.

but they did not think it sounded right to say

(18) The chairman helped the environment in order to increase profits.

In other words, people were willing to apply reason explanations to the morally bad behavior but not to the morally good one (Knobe 2004).

Note that the asymmetry here does not apply to explanations of all types; it applies only to reason explanations. Thus, in both cases, it is clear that the chairman performed the behavior because he had a desire to increase profits. The difference between the cases emerges only when we ask whether or not he performed this behavior for a reason. In the case of the morally bad behavior, people do say that he acted for a reason and that his reason had something to do with increasing profits. In the case of the morally good behavior, by contrast, people seem to think that reason explanations are inappropriate. Faced with this latter case, many people say: ‘He didn’t actually help the environment for a reason. Rather, he implemented the policy for a reason (namely, to increase profits), and his implementing of the policy just happened to involve helping the environment as well.’
Looking at these results for the first time, one can easily come to feel that there must be some kind of trick involved. It seems obvious somehow that the acceptability of a reason explanation just can’t have anything to do with the moral status of the behavior to be explained. To the extent that people’s responses seem to be influenced in some way by moral considerations, one may conclude that the effect must be due to an experimental artifact or some similar form of error. Let us therefore pause for a moment to consider a number of possible objections.

To begin with, I should emphasize that none of the objections I will be discussing have actually been put forward in the existing philosophical literature. Philosophers have offered competing explanations of the findings described above, but all of the philosophers who have discussed these findings have concluded that people’s use of reason explanations is indeed sensitive in some way to moral considerations (Adams & Steadman 2004b; Knobe forthcoming; Nadelhoffer 2005; Turner forthcoming). In other words, nothing in what I have said so far is actually controversial among philosophers working on these issues. Still, it may be helpful to pause briefly here to allay some lingering doubts.

Let us turn now to the first potential objection:

In the cases under discussion, one finds three basic elements — a behavior, an aim and a side-effect. These three elements can be depicted as follows:

Now, the phrase ‘in order to’ is used to indicate a means-end relationship. In other words, it is used to indicate the kind of relationship that can only obtain between a behavior and an aim.
Strictly speaking, then, the phrase cannot be applied to the bringing about of a side-effect. So what people really mean to say is not that ‘the agent brought about the side-effect in order to bring about the aim’ but rather that ‘the agent performed a behavior [which happens to lead to the side-effect] in order to bring about the aim.’ To the extent that people sometimes fail to express this point with sufficient precision, it is only because they are being sloppy. Nothing of importance is thereby revealed about the nature of reason explanation itself.

The key claim behind this objection is that the phrase ‘in order to’ can only be correctly applied in cases that meet a specific criterion. But what the data show is that people sometimes use that phrase in cases that do not meet this criterion. So the objection posits a standard of ‘correctness’ that departs in certain respects from ordinary usage.

One way to make this objection compelling would be to show that people find it difficult to keep track of all the complex conceptual distinctions that seem to be in play here. For example, it could be argued that people have a hard time getting a handle on the distinction between saying ‘The agent brought about the side-effect in order to…’ and saying ‘The agent performed a behavior [which happens to lead to the side-effect] in order to...’ But the data seem to suggest that people actually are capable of using this distinction. In particular, people seem not to be willing to say that the agent brought about a side-effect for a reason when that side-effect is a good one. The objection therefore faces a major hurdle. It seems unlikely that people are able to make all of the necessary conceptual distinctions as long as the side-effect is a good one but that they somehow find these distinctions too difficult when the side-effect is a bad one. A more plausible explanation would be that they make all the necessary conceptual distinctions in both cases but that, when the side-effect is a bad one, they regard reason explanations as acceptable.

Even so, there are various strategies one might use to argue for a standard of ‘correct’ use that diverges in certain respects from the use one finds among ordinary speakers. It must be recalled, however, that we are not concerned here with questions about the correct use of reason explanations; our concern lies rather with the ordinary folk-psychological use of reason explanations. In particular, we were concerned with the question as to whether the ordinary folk-psychological practice makes any use of moral
considerations. Additional questions may arise as to whether this folk-psychological practice is correct or incorrect, but those questions lie outside the scope of the present chapter.

We can therefore turn to a second objection:

The sentence ‘He harmed the environment in order to increase profits’ involves an important ambiguity.

On one interpretation, the sentence means something like ‘There exists a behavior that is a harming of the environment, and he performed this behavior in order to increase profits.’ On this interpretation, the sentence is true. After all, the behavior of harming the environment simply is the behavior of implementing the policy. (What we have here are two descriptions of the very same behavior.) Moreover, it is clear that he implemented the policy in order to increase profits. So there is a sense in which the behavior of harming the environment was a behavior performed in order to increase profits.

But there is also an interpretation on which the sentence comes out false. On this latter interpretation, the sentence says that the behavior was specifically performed for a reason under the description ‘harming the environment.’ But this is clearly not the case. The behavior was not performed for a reason under that description but rather under the description ‘implementing the policy.’

In short, it can be correct to regard a sentence like this one either as true or as false, depending on which interpretation one happens to adopt.

This objection takes us into difficult territory in the theory of action. Some theories of action do indeed imply that the harming of the environment and the implementing of the policy are the same behavior under two different descriptions (Anscombe 1957; Davidson 1967), while others imply that what we have here are two distinct behaviors that just happen to involve the same bodily motions (Goldman 1970). I cannot hope to resolve this issue here.

Fortunately, however, it is possible to address the central questions of the present chapter without taking any stand on the difficult issue of action individuation. In fact, it
seems to me that the whole issue is a red herring in the present context. What needs to be explained is the asymmetry between people’s judgments regarding good side-effects and their judgments regarding bad side-effects. The alleged ambiguity here seems not to provide an explanation of that phenomenon unless one assumes that there is a systematic pattern whereby people resolve the ambiguity differently depending on the goodness or badness of the side-effect. I see no grounds for positing such a systematic pattern.

In the remainder of the paper, I therefore ignore these objections and proceed on the view that has become the consensus among researchers working on this issue. That is to say, I will proceed on the view that a certain kind of normative judgment — a judgment about the goodness or badness of a side-effect — really does play a role in people’s use of reason explanations.

VI

What we need to know now is why moral considerations play this role in the practice of reason explanation. Note that this is not the sort of question that can be answered by simply describing the conditions under which people deem reason explanations acceptable. Even if we had a precise list of necessary and sufficient conditions, there would still be a legitimate question as to why people followed this particular list rather than some other one.

One obvious strategy for answering this question would be to consider the purposes that the practice of reason explanations serves in people’s lives. We could try to show that these purposes are better served by a practice that is in some way sensitive to moral considerations than by a practice that leaves moral considerations out of account. This is the strategy we used in the previous chapter, where we tried to show that certain features of the concept intentional action could be understood in terms of the purposes that this concept serves in people’s lives. It is also a natural approach to understanding people’s use of concepts more generally. If we wanted to understand the peculiar features of, e.g., the concept knowledge, a natural place to start would be by thinking about what people ordinarily do with this concept and trying to show that they wouldn’t be able to do quite as good a job of it if the concept lacked the features in question.
The trouble is, the asymmetry we have uncovered does not seem to contribute to any of the purposes to which reason explanations are normally put. Indeed, the more one considers the fundamental purposes that reason explanations serve in people’s lives, the more baffling the asymmetry comes to seem. It seems that people use reason explanations for broadly *scientific* purposes (where they are concerned with prediction, explanation and control), as well as for more *normative* purposes (where they are concerned with questions about whether a behavior was truly justified), and perhaps reason explanations are also used for various other purposes. But none of these purposes seems to be furthered in any way by the moral asymmetry we have been discussing. If people simply accepted reason explanations for all side-effects, it seems that we could still achieve all of these purposes perfectly well.

At this point, one might be tempted to conclude that the asymmetry is entirely pointless, just a bizarre quirk of our cognitive capacities. But it seems to me that this conclusion is a bit too hasty. Although the asymmetry itself may not directly help us to achieve any important purposes, I want to suggest that there is a more indirect way in which it can be seen as useful. The key thing to keep in mind is that the asymmetry we observe is produced by the interaction of certain underlying cognitive mechanisms. These mechanisms might not be used exclusively for thinking about reason explanations; perhaps they are also used for various other purposes. Hence, it might be possible to show that it actually makes sense for the underlying mechanisms to work the way they do (because of their importance in achieving other purposes) and that they end up creating an asymmetry in the use of reason explanations as a kind of by-product.

To get a rough sense for what I am after here, it might be helpful to look at a more mundane example. Consider the person who lights a fire by using some newspaper as kindling. Someone might point out that the writing on the newspaper isn’t playing any useful role at all in the process of lighting the fire. But one should not therefore conclude that the writing is entirely pointless, just a bizarre quirk of the kindling this person happened to use. The writing does play a helpful role in *another* purpose for which newspapers are used, a purpose that has nothing to do with lighting fires. Then, given that the newspaper was already around for this other purpose, people came to use it for kindling as well.
What I want to suggest now is that the use of moral considerations in reason explanations can be understood in a similar way. There is no need to suppose that moral considerations actually play any useful role in the activity of reason explanation. Rather, what we need to do here is look at other activities that use the same cognitive resources. The hypothesis will be that moral considerations actually do play a useful role in these other activities and that our basic cognitive resources have developed accordingly. Then, given that we already had these cognitive resources around for use in another activity, we began using them for reason explanations as well.

VII

The relation I wish to explore is a familiar one. As numerous authors have noted, there seems to be an important connection between the practice of reason explanation and the concept of intentional action (Anscombe 1957; Goldman 1970; Malle et al. 2000; Mele 1992). Philosophers continue to disagree about the precise nature of this connection, but most would agree with the basic claim that reason explanations cannot properly be applied to behaviors that were not performed intentionally.

The standard view here — first introduced by Anscombe (1957) — is that, when we come to a correct understanding of the practice of reason explanation, we will thereby arrive at a more fundamental understanding of the concept of intentional action. This standard view starts out with the assumption that we have some independent notion of what it means for a behavior to be performed for a reason and that we can then use this notion to arrive at an understanding of the distinction between intentional and unintentional behavior. So, for example, Anscombe suggests that there is a peculiar sense of the question ‘why?’ in which this question must be understood as a request for a reason explanation. One can then define the word ‘intentional’ by saying that an intentional action is a behavior about which this sort of question can be legitimately asked.

Note that this account makes no reference to the ways in which the concept of intentional action figures in the process of assigning praise and blame. The assumption is that we can define ‘intentional’ without making any mention of moral considerations.
Then, once we have this definition in hand, we can use the concept of intentional action in certain kinds of moral judgments.

It seems to me that this account cannot make sense of the data we have amassed thus far. We have seen that moral considerations play a role both in our practice of reason explanation and in the concept of intentional action. More importantly, we have seen that moral considerations play the same role in both of these cases — bad side-effects are regarded as intentional and can be explained using reasons, whereas good side-effects are not regarded as intentional and cannot be explained using reasons. So far, none of this creates any difficulties for the standard view. But now we arrive at the problem. There seems to be an asymmetry here between the concept of intentional action and the practice of reason explanation. In the context of intentional action, it makes sense that moral considerations would play a role. As we saw in the previous chapter, the concept of intentional action is used in making judgments of praise and blame, and the influence of moral considerations makes the concept more useful in this task. By contrast, moral considerations do not seem to play any useful role in our practice of reason explanation. That is to say, when we consider the purposes for which people ordinarily use reason explanations, it doesn’t appear that a sensitivity to moral considerations would allow us to do a better job of serving those functions. On some level, the use of moral considerations in reason explanations seems pointless and therefore a bit baffling.

In light of these considerations, it seems more than a little perverse to suggest that we have some prior understanding of what is involved in acting for a reason and that we can then just define the concept of intentional action in terms of that prior understanding. Surely, it would be more plausible to suggest that the process goes in the opposite direction. We start out with a concept of intentional action (a concept in which moral consideration played a key role). Then, given that we need this concept anyway, we come to use it in assessing reason explanations as well.

Perhaps the basic idea behind this hypothesis is best conveyed by means of a fable. Like most fables, this one is not intended to be literally true, but I hope that it does help to illustrate certain important aspects of the processes involved here. Our fable begins with a group of people trying to construct a concept that can be used, above all, in the assignment of praise and blame. Faced with this task, they create a concept which
they call ‘the concept of intentional action.’ But, some years later, a new problem arises. The people are developing a practice of reason explanation, and they need a concept that they can use to pick out the class of behaviors for which reason explanations are appropriate. One person says: ‘What we need now is a new concept — one that is perfectly suited to the task of picking out behaviors for which reason explanations are appropriate.’ But another person interrupts: ‘Wait! We already have the concept of intentional action, and although that concept isn’t perfectly suited to the task at hand, it would certainly do a fairly good job. So instead of creating a whole new concept, maybe we should just try to do as well as we can with the one we’ve already got in place.’ In the end, the people choose not to construct any new concepts. Perhaps they modify the concept of intentional action in some minor respects to make it workable in their practice of reason explanation, but the main features of the original concept remain exactly as they had been. The influence of moral considerations — which seems to make sense when the concept of intentional action was being used primarily in judgments of praise and blame — therefore ends up appearing in people’s reason explanations as well.

Clearly, certain aspects of this fable are not intended to be understood literally. The precise chronological order of the events, for example, is of no importance. In other words, it is of no importance whether people first developed the concept of intentional action and then developed the practice of reason explanation. These events could have occurred in the opposite order, or — what seems more likely — they could have occurred simultaneously. What is important here is not the chronological order but the order of explanation. The key claims are that (a) moral considerations play a role in the practice of reason explanation because they play a role in the concept of intentional action and that (b) moral considerations play a role in the concept of intentional action because the role of moral considerations in this concept enables us to do a better job of assigning praise and blame. The previous chapter argued for the second of these claims; we turn now to the first.
Before turning to the evidence for and against this claim, we need to get clear about precisely what sort of claim it is. Above all, it should be emphasized that it is *not* a claim about the overall pattern of people’s intuitions. That is, it is not a claim that we can capture the pattern of people’s intuitions about reason explanations by offering an analysis that treats intentional action as a necessary condition. The aim here is not to come up with a set of criteria that ‘fit’ people’s intuitions but rather to uncover the criteria that *people actually use* when they are trying to determine whether or not a reason explanation is acceptable.

Hence, the claim is not just that the behaviors people regard as intentional turn out to be the very same behaviors that people explain using reasons. Rather, the claim is that people themselves actually evaluate reason explanations by using their concept of intentional action. This is not the sort of claim that can be made true by certain patterns in people’s intuitions. At best, people’s intuitions serve as a form of evidence as to whether the claim is true or false. As it happens, though, the main source of evidence for the claim lies in the fact that it provides such a simple and elegant explanation of the intuitions we have been discussing thus far.

I must confess, however, that recent experimental results have uncovered intuitions that seem not to accord very well with the hypothesis I have proposed. In particular, there appear to be behaviors that people explain using reasons but which people do not regard as intentional. The only way to accommodate these new results on the theory I have proposed is to introduce certain ad hoc auxiliary hypotheses. Clearly, the fact that I am forced to introduce ad hoc hypotheses here provides at least some reason to reject my theory. The key question now is whether it is possible to construct an alternative theory that provides a simple and unified explanation of all of the data, including not only the data presented above but also the data I present in this section. (I have been unable to come up with such an alternative theory, but perhaps others will have greater success.)

My one consolation here is that the data I am about to present spell trouble not only for my own theory but also for all existing accounts of the relationship between
intentional action and reason explanation. All of these accounts seem to suggest that people will not offer reason explanations for behaviors that they do not regard as intentional. The difference is just that, where previous accounts tried to define ‘intentional’ in terms of an independent notion of acting for a reason, I suggest that people have an independent concept of intentional action and then use it to evaluate reason explanations.

(1) That said, I turn to the first problem. This problem arises in cases where an agent is trying to perform a behavior and actually does perform that behavior but only succeeds in performing the behavior through sheer luck. For a particularly extreme example, consider a person who needs some extra money and therefore decides to play the lottery. He guesses five numbers at random, and — in an extraordinary burst of luck — he ends up guessing all five correctly. So he actually does win the lottery, exactly as he had hoped. Faced with this case, most people do not find it acceptable to use either the sentence ‘He won the lottery intentionally’ (Knobe 2005a) or the sentence ‘He won the lottery in order to have more money’ (Knobe unpublished data). The issue here is that the agent seems not to have enough control over the outcome; his success owes too much to luck.

It is difficult to say precisely how people know when an outcome involves ‘too much luck,’ and I don’t think that anyone has yet been able to formulate a persuasive answer to this question. (Certainly, it is not just a matter of the agent’s probability of success.) But our aim here is not to figure out precisely which rule people are using. The question is simply whether people use the same rule in both cases — whether they use the same rule to determine whether a behavior was performed intentionally as they use to determine whether a behavior was performed for a reason. Suppose people determine (in whatever way this is done) that a particular behavior owes too much to chance and therefore was not performed intentionally. If people actually use their concept of intentional action to evaluate reason explanations, they should then reject all reason explanations.

---

4 For example, consider a safe that works on more or less the same principle as the lottery we described above. The safe only opens when a person correctly types in five numbers. Now suppose that an agent has no idea what the right numbers are and therefore decides to guess at random. The agent guesses correctly, and the safe opens. Here it sounds far more correct to use sentences like ‘He opened the safe intentionally’ or ‘He opened the safe in order to get the money.’ (I am not at all sure what difference between this case and the lottery case above accounts for the difference in people’s intuitions.)
explanations for this behavior. In other words, any behavior that involves too much luck to be regarded as intentional should also be a behavior that involves too much luck to be explained using reasons.

But this is not the pattern that we actually observe. It is possible to construct a case in which people feel that there was too much luck for the behavior to count as intentional but not too much luck for the behavior to be explained using reasons. In fact, this surprising pattern of responses can be found in the ‘rifle contest’ case presented in the previous chapter. Jake realizes that he can only win the contest if he hits the bull’s-eye. He aims at the bull’s-eye and presses the trigger… but he is an incredibly bad rifleman. The gun slips in his hand and the shot goes wild. Nonetheless — through sheer luck — he ends up hitting the bull’s-eye. Faced with this case, most people do not feel that it is correct to say ‘Jake hit the bull’s-eye intentionally’ (Knobe 2003) but they do feel that it is correct to say ‘Jake hit the bull’s-eye in order to win the contest’ (Knobe unpublished data). Thus, there do appear to be cases in which people are not willing to apply the concept of intentional action but in which they are willing to use reason explanations.

One response to this difficulty would be to conclude that the practice of reason explanation and the concept of intentional action are, in fact, entirely unrelated. But this sort of view fails to account for a number of important generalizations. Even in the present context, we see a remarkable similarity between the conditions under which people use reason explanations and the conditions under which they apply the concept of intentional action. In both cases, we find that people’s intuitions are sensitive to the amount of luck involved; the difference lies only in the precise cut-off point at which people determine that the amount of luck is too great. It therefore seems preferable to find some way of reconciling the occasional divergences in the use of the two concepts with the view that they are nonetheless intimately related.

Perhaps it would be helpful here to take a look at other concepts that appear to be related. Take the concept friendship. The obvious way to explain this concept would be to provide a list of features — friends like each other, friends engage in social interactions with each other, friends help each other, and so on. But it seems that a relationship can be considered a friendship even when it does not have all of these features. Consider the
feature *engaging in social interactions with each other*. It seems that this feature specifies, however vaguely, a certain segment of a continuum. The continuum goes from people who have never interacted in any way to people who spend every moment of every day interacting socially with each other. People on one side of the continuum clearly do not engage in social interactions; people on the other side clearly do engage in social interactions; and then there are borderline cases in the middle where it isn’t quite clear what we should say. Now, consider a pair of people who fall just below the crucial threshold, so that we would be inclined on the whole not to say they ‘engage in social interactions.’ If these people showed all of the other relevant features of friendship to a particularly high degree, we would nonetheless say that they were ‘friends.’ In other words, even though there is an important connection between the concept of friendship and the feature of engaging in social interactions, we will sometimes be inclined to regard two people as friends even when they fall below the threshold necessary for having that feature.

One common way of conceptualizing this phenomenon is to suppose that each feature is assigned a ‘weight.’ For each feature that the object has, it gets a certain number of points corresponding to the weight of that feature. If the number of points goes above a certain critical number, the object is said to fall under the relevant concept (Rosch 1978; Smith & Medin 1981). But in cases like the ones we are faced with here, it seems wrong to suppose that people simply make yes/no judgments regarding each feature. A more plausible view would be that people sometimes conclude that an object has a feature to a particular degree. In such cases, one cannot add the whole weight normally assigned to the feature, but it is not as though the borderline presence of the feature counts for nothing at all.

Quite possibly, the pattern we observe in people’s intuitions about reason explanation and intentional action can be explained along similar lines. Although people feel that Jake’s success involved too much luck to count as fully intentional, they do feel that it was intentional to at least some small degree. This borderline judgment of intentional action then feeds in to the process by which people evaluate reason explanations, and they end up concluding that the behavior falls above the critical threshold to be considered eligible.
(2) A more difficult problem arises when we begin to think about precisely what it means for a given side-effect to be ‘bad.’ One wants to know whether people’s intuitions are actually shaped by judgments about whether the side-effect truly is bad or whether their intuitions are only shaped by judgments about whether the agent thinks the side-effect is bad.

Perhaps the best way to get at this question is to construct a case in which the side-effect seems truly to be a good one but in which the agent somehow comes to think that it is a bad one. Here is one such case:

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.

This case gives us a chance to see precisely which moral judgment is influencing people’s intuitions. If people’s intuitions are shaped by their judgment about what truly is bad, they should say that the saving of the Americans was not intentional and was not done for a reason. By contrast, if people’s intuitions are shaped by their judgment about what the agent thinks is bad, they should say that the saving of the Americans was intentional and done for a reason.

The actual result, however, is surprising and puzzling in a way that neither of these options would have been. People say that the saving of the Americans was not
intentional but that it was done for a reason (Knobe & Kelly 2006). This result suggests that people’s intuitions about intentional action are shaped by their judgments about whether the side-effect truly is bad but that their intuitions about reason explanations are shaped by their judgments about whether the agent thinks the side-effect is bad. In short, we have on our hands a second respect in which the criteria for reason explanation come apart from the criteria for intentional action.

Here again, one possible reaction would be to conclude that the practice of reason explanation does not actually have anything to do with the concept of intentional action. Perhaps there are just two distinct mental processes here — one that uses people’s judgments about what truly is bad, another that uses people’s judgments about what the agent thinks is bad. It is certainly conceivable that this hypothesis will turn out to be correct, but it seems that we do have at least some reason to prefer an alternative explanation.

One initial difficulty comes from the fact that people’s intuitions about reason explanations do not seem to be affected only by their judgments about what the agent believes to be bad. In the case of the executive harming the environment, for example, there is no indication at all that the agent regards the environmental harm as a bad thing. (In fact, he says explicitly that he does not care at all about the environment.) Yet people do seem to be classifying the side-effect as ‘bad’ in some relevant sense and therefore applying reason explanations to the behavior.

So perhaps the most natural way to interpret the data would be to say that people’s intuitions are shaped both by their judgments about what the agent thinks is bad and by their judgments about what truly is bad. If the wording of the vignette makes it extremely salient that the agent regards the side-effect as bad, then people will permit reason explanations regardless of whether they themselves think that the side-effect truly is bad. But that is not to say that people’s beliefs about whether the side-effect truly is bad...

---

5 Subjects were asked either whether it was right to say ‘The terrorist saved the Americans in order to rescue his son’ or whether it was right to say that ‘The terrorist saved the Americans intentionally.’ Most subjects said that the first sentence was right, whereas relatively few subjects said that the second sentence was right.
bad are completely irrelevant. They do have an impact in cases where the agent’s own beliefs are not made especially salient.

When we begin to think about reason explanations in this way, it seems that people’s intuitions about reason explanations are actually quite similar to their intuitions about intentional action. After all, recent studies have shown that people’s intuitions about intentional action show at least some sensitivity to their judgments about what the agent regards as bad; it’s just that these intuitions are far more sensitive to people’s judgments about what truly is bad (Knobe & Kelly 2006). In other words, changes in people’s judgments about what truly is bad will have a large impact on their judgments about whether behavior was performed intentionally, but it is not as though changes in people’s judgments about what the agent regards as bad have no impact at all. They do have an impact, just a smaller one.

In short, it seems that the difference between people’s intuitions about reason explanations and their intuitions about intentional action is simply a matter of degree. Both sorts of intuitions are affected in part by judgments about what the agent thinks is bad and in part by judgments about what truly is bad. The only difference lies in the relative weights. Intuitions about reason explanations are more sensitive to judgments about what the agent regards as bad, whereas intuitions about intentional actions are more sensitive to judgments about what truly is bad.

Now that we have a rough sense for the overall pattern of people’s intuitions, we can return to the basic question that these experiments were designed to address. That question, recall, was whether (1) people’s intuitions about intentional action actually serve as input to the process that generates intuitions about intentional action or (2) these two kinds of intuitions are generated by two completely separate mechanisms. When one considers the pattern revealed in recent experiments, it seems that strong doubt is cast on the second of these hypotheses. The problem here is that the criteria underlying judgments about reason explanations and the criteria underlying judgments about intentional action are, despite the evident differences, extremely similar in a variety of important respects. If we simply assume that there are two entirely separate mechanisms here, we will be left without any explanation of these similarities.
A more promising explanatory strategy, then, would be to suggest that both types of intuition are produced by the same underlying mechanism but that this mechanism can generate slightly different results depending on the particular question it is faced with. When the question is about intentional action, more weight is assigned to the feature of truly being bad. When the question is about reason explanation, more weight is assigned to the feature of being regarded as bad by the agent. But either way, the underlying mechanism is the same one.

IX

When we examine people’s intuitions about reason explanations, we find a puzzling pattern. People’s intuitions appear to be governed by rules that do not help them — at least in any obvious way — to achieve the purposes we normally associate with the practice of reason explanation. This phenomenon comes as a surprise, and it is hard to know quite what to make of it.

My aim here has been to suggest a way in which moral considerations might end up playing a role in people’s intuitions about reason explanation. Specifically, I suggested that moral considerations play a role in the concept of intentional action and that they can therefore have an impact (albeit indirectly) in intuitions about reason explanation. Further research will be needed before we can know whether this hypothesis is on the right track, but if it does turn out to be correct, it will open up a new approach to certain difficult questions in the study of folk psychology. It will be very interesting to see whether the basic strategy explored here only seems plausible when applied to reason explanations or whether similar accounts can help to illuminate the puzzling phenomena we observe in the use of other folk-psychological concepts.
Causal Attributions

When we are trying to determine whether or not an agent deserves blame for an outcome, it seems that we first need to answer two kinds of questions:

1. Questions about the status of the outcome itself — whether the outcome is good or bad, how much of an impact it will have, etc.

2. Questions about whether the agent stood to that outcome in the right sort of relation — whether the agent caused the outcome, whether she brought it about intentionally, etc.

The standard view among social psychologists is that these two kinds of questions are of two very different types. The first kind of question is taken to be normative; the second is taken to be purely descriptive.

In this respect at least, social psychology is in agreement with common sense. Questions about the goodness or badness of a given outcome are quite clearly normative questions. By contrast, questions of the second kind do appear to be purely descriptive. Thus, causation seems to be a purely naturalistic relation between events, and the distinction between intentional and unintentional behaviors seems to be a straightforwardly psychological one.

But perhaps things are not quite as simple as they appear. In the chapter 2, we saw that moral considerations actually play a key role in people’s judgments as to whether or not a behavior was performed intentionally. A question now arises as to whether or not it might be possible to find a similar effect in people’s causal attributions.

Here I will argue that such an effect does indeed obtain. The claim will be that causal attributions are not purely descriptive judgments. Rather, people’s willingness to say that a given behavior caused a given outcome depends in part on whether they regard the behavior in question as morally wrong. To a first glance, this position may seem implausible, even a bit bizarre, but perhaps it will gain some plausibility when we consider people’s intuitions about certain kinds of cases.
Before considering those intuitions, however, we need to make a few initial clarifications. First, it is essential to get clear about what social psychologists mean by ‘causal attribution.’ Then we need to distinguish our own (more philosophical) inquiry from two other projects with which it might easily be confused.

We begin, then, with the basic idea of causal attribution. In a typical attribution experiment, subjects are not merely asked to make yes/no judgments about whether or not a given factor counts as a cause. Rather, they are asked about the degree to which the factor is a cause. Thus, subjects might be asked to choose a point on a scale from 0 to 7, with 0 marked ‘not at all a cause’ and 7 marked ‘very much a cause.’ A central goal of attribution theory is to understand why people might attribute an outcome more to some factors and less to others.

This conception of causal attribution as a matter of degree seems to accord with our ordinary linguistic practice. In an ordinary conversation, one might say: ‘The accident was due primarily to the bad road conditions but also partly to the supervisor’s negligence.’ Here both the road conditions and the negligence are declared to be causes, but the speaker attributes the outcome more to the road conditions than to the negligence. And, of course, it might make a great deal of difference how much the outcome is attributed to each of these factors. For example, the degree to which we blame the supervisor for the accident might depend in part on the degree to which we attribute it to his behavior.

Nonetheless, the focus of most philosophical theories of causation has been on the prior (and absolutely essential) question as to whether a given factor can be considered a cause at all. Although philosophical theories of causation differ in a number of important respects, most of them do not attempt to answer the question as to when it is right to attribute an outcome more to one factor than to another. Indeed, the account developed here will be more or less neutral between these competing philosophical theories.

Still, there do seem to be at least two lines of existing research that treat attribution as a matter of degree. We can get clear about the nature of our own inquiry by distinguishing it more carefully from each of these other lines of research.
First of all, it is important to distinguish the topic of our inquiry from a seemingly related issue in the philosophy of science. There, a question arises about how scientists ought to determine the degree to which a given factor contributed to a given outcome. The aim is construct a concept of ‘causal strength’ that might serve a useful role in scientific investigations.

So, for example, Northcott (2004) suggests that we define the degree to which factor \( f \) influenced outcome \( o \) relative to the contrast case \( f_1 \) and the background conditions \( w \) as \( P(o \mid f \& w) - P(o \mid f_1 \& w) \). Thus the suggestion is that, if scientists wanted to assess the degree to which poor infant nutrition contributed to a given woman’s later health problems, they should take the probability of that woman having health problems given that she had poor infant nutrition and then subtract the probability that she would have health problems if she had not had poor infant nutrition. Arguing against all theories of this general type, Sober (1988) suggests that there is no need to construct a single unified concept of causal strength that can be applied to all possible sciences. As he puts it: ‘there is no such thing as the way science apportions causal responsibility; rather, we must see how different sciences understand this problem differently’ (1998; p. 304).

This issue seems to me to be an important one, and I have no objections to those philosophers of science who are actively working on it today. Still, it must be emphasized that our aim here is quite different. We are not at all interested in constructing a concept that will prove useful in the sciences. Our aim is simply to understand what people are doing when, under ordinary circumstances, they make an attribution. Of course, the principles governing people’s ordinary attributions might turn out to be well suited to the systematic sciences, and in that case, our inquiry might eventually arrive at a kind of convergence with the parallel inquiry in the philosophy of science. But this sort of convergence is in no way guaranteed. It is surely possible that certain aspects of people’s ordinary attributions will turn out to be perfectly useless, perhaps even downright harmful, when applied in the systematic sciences. But no matter. Our aim is not to construct a concept that has certain desirable properties but rather to understand what people are actually doing when they make an attribution.
At the same time, however, our inquiry must be distinguished from the project within social psychology of developing theories that predict which attributions people will make under particular circumstances. The aim of this project is to understand all of the many psychological forces that can influence people’s attributions — including biases due to salience, limited capacity, self-serving motivations, and so forth (e.g., Jones & Davis 1972; Nisbett & Ross 1980; Taylor & Fiske 1978).

Here, however, we will not be concerned with errors and biases. Instead, our concern lies with the more fundamental question as to what people are trying to do when they make causal attributions. Putting this point another way, one might say that our aim is to figure out precisely what question people’s causal attributions are supposed to be answering.

In a certain sense, then, our inquiry might be compared to the sort of thing one finds in linguistics. It is well known that people make linguistic errors when they are distracted, intoxicated or emotionally aroused. But the typical linguist is not interested in understanding all of the factors that influence people’s utterances. The goal is rather to develop a model of the fundamental competence underlying people’s linguistic behavior. To achieve this goal, one begins with observations of people’s linguistic behavior, but the aim is not merely to catalog regularities in that behavior. Rather, one hopes to use the behavior as a basis for inferences regarding the competence that generated it.

In a similar way, as we investigate the competence underlying people’s attributions, we will be guided primarily by evidence concerning the attributions that people actually make. But our aim is not simply to find certain regularities in the observed attributions. The goal is rather to develop a theory of the fundamental competencies that generated them.

With these distinctions in place, we can now explain more clearly what we mean by the claim that normative considerations play an important role in attribution. On one hand, we definitely aren’t in the business of constructing a new concept that has certain desirable characteristics. But on the other hand, we aren’t simply making a claim about observable regularities in the attributions people actually make. The claim is that normative considerations play a fundamental role in the competencies underlying
people’s attributions. We can test this claim by looking at people’s actual attributions and then developing a theory that explains them.

II

Let us begin, then, by examining a few cases of attributions that do seem to suggest a role for normative considerations.

First, take cases in which people attribute an outcome to an omission, i.e., to the fact that someone did not perform a particular action. This is the sort of attribution that we might express with a sentence like: ‘The main reason why the child died is that his parents did not give him enough food.’ Here the attribution is not to anything that the parents actually did but rather to the fact that they didn’t do something — namely, give the child enough food.

It can be quite difficult, in such cases, to say exactly when it is correct to attribute a given outcome primarily to a given omission. Typically, there will be a great number of people who could (at least in principle) have performed an action that would have prevented the outcome. Thus, in our case of the malnourished child, we immediately attribute the child’s death to the parent’s failure to provide enough food. But surely there are also other people who could have provided the child with food. The lawyer down the street could have quit his job, formed an organization to help malnourished children, and thereby provided enough food to save the child’s life. And yet, we would never attribute the child’s death primarily to the lawyer’s failure to form such an organization. Instead, we attribute it primarily to the parent’s failure to feed him.

What exactly is the difference here between the parents and the lawyer? As Thomson (2003) and McGrath (2004) have argued, people’s intuitions in cases like these can be influenced by the normative status of the omissions themselves. Thus, part of the difference between our two cases is that the parents specifically ought to have provided enough food for the child. It is a difference that comes down, not merely to certain descriptive facts, but also to questions of right and wrong, of good and bad, of people’s responsibilities to each other.

Now, one problem with our case of the dying child is that there are too many uncontrolled variables. Let us assume, as seems plausible, that people would attribute the
child’s death to his parents’ failure to feed him rather than to the lawyer’s failure to form a charitable organization. Still, there are a wide variety of respects in which the position of the parents differs from the position of the lawyer. How are we to know whether the relevant difference — the difference that leads us to attribute to one factor rather than another — lies in the normative considerations? Ideally, we would want to find two factors that were exactly alike in all of their descriptive properties and differed only in their normative properties. Then we could be certain that any differences in people’s attributions were due to normative differences between the factors and not to purely descriptive differences. Unfortunately, however, this ideal kind of experimental design is blocked by the supervenience of the normative on the descriptive. (Two factors cannot differ in any of their normative properties unless these factors also differ in some of their descriptive properties.)

Still, we can try to find factors that differ from each other in more limited and well-understood ways than the factors in our child’s death story do. The impact on attributions will then be more subtle, but we will also have less difficulty in locating the source of that impact. Consider, for example, the following vignette:

George and Harry both work in a large office building. George is the janitor; Harry takes care of the mail.

Every day, George goes through the entire building and empties all of the garbage baskets. Since the building is large, this task normally takes him about one half hour.

One day, George is feeling tired and decides not to take out the garbage.

Harry sees that the garbage hasn’t been taken out. He doesn’t go to take it out himself, since that is not his job.

The next morning, all of the trash is still sitting in the garbage baskets. A bad odor fills the building.

Here people would tend to attribute the odor more to George’s failure to take out the trash than to Harry’s failure to take out the trash. But why? Both George and Harry were capable of taking out the trash; both chose not to do so; and if either had taken out the trash, the odor would not have filled the building. It seems to me that the most plausible explanation here is the one that appeals to normative considerations. The reason why we
attribute the odor more to George’s failure to take out the trash is that George really
*ought* to have taken out the trash. Thus, it is a normative consideration—a consideration
related to what George ought to have done—that is leading us to attribute more to one
factor than to another.

This point is not related specifically to cases of omission. It applies with equal
force to cases in which an agent actually performs an action and we are wondering
whether or not that action led to the outcome. For example, consider the following
vignette:

Lauren and Jane work for the same company. They each need to use a computer for
work sometimes.

Unfortunately, the computer isn’t very powerful. If two people are logged on at the
same time, it usually crashes.

So the company decided to institute an official policy. It declared that Lauren would
be the only one permitted to use the computer in the mornings and that Jane would be
the only one permitted to use the computer in the afternoons.

As expected, Lauren logged on the computer the next day at 9:00 am.

But Jane decided to disobey the official policy. She also logged on at 9:00 am.

The computer crashed immediately.

Here we would attribute the crash more to Lauren’s behavior than to Jane’s behavior. But
why? The two factors appear to be similar in many ways. The most plausible hypothesis
seems to be that this difference in people’s attributions stems from a difference in
perceived normative status—i.e., from a sense that Jane was not doing anything wrong
but Lauren really ought to have refrained from using the computer.

**III**

Our chief aim here is simply to present a convincing case for the conclusion that
normative considerations play *some* role in attribution. But to make this case truly
convincing, we need to present a more specific hypothesis about what role it is that normative considerations are playing.

The problem is that, as soon as one tries to clarify the role of normative considerations in people’s attributions, one finds oneself confronted with more general questions about how attributions work and when they can be considered correct. In the present context, however, I will be doing my best to avoid these broader questions. Since my sole aim is to provide an account of the role of normative factors, I will try to remain neutral about many of the key questions regarding the overall nature of attributions. The basic strategy will be to give an extremely cursory and handwaving account of causal attribution as a whole and then explore in somewhat more detail the role that normative considerations might play in that account. To the greatest extent possible, we want to avoid relying on any contentious claims about the nature of causal attribution as a whole and hence to assure ourselves of being compatible with the correct overall account of causal attribution, whatever that account turns out to be.

I begin with the assumption that an outcome is always attributed to a factor relative to a particular contrast case. In ordinary conversations, the contrast case is rarely stated explicitly. It is usually given either by the conversational context or by the speaker’s intonation. Thus, compare the following sentences:

— “The party was so boring because Dave arrived at 10 pm.”
— “The party was so boring because Dave arrived at 10 pm.”

Although these two sentences attribute the boringness of the party to the same factor (Dave arriving at 10 pm), they differ markedly in their contrast cases. The first sentence contrasts this factor with the case in which Dave arrives at a time other than 10 pm. In essence, it says that the party was boring because Dave arrived at 10 pm rather than arriving at a different time. The second sentence contrasts the very same factor with a different case — namely, the case in which someone other than Dave arrived at 10 pm. It seems clear that one of these sentences can be true even if the other is false.

The key to understanding people’s attributions is to understand the relation in which an outcome must stand to a given factor and its contrast case before the outcome
can be attributed to the factor. An obvious first stab at an explication of this relation would be to say that an outcome is correctly attributed to a given factor if and only if:

1. the outcome actually occurred,

2. the factor actually arose,

3. if instead the contrast case had arisen, the outcome would not have occurred.

But it can easily be shown that this analysis breaks down in certain cases (e.g., in cases of overdetermination). Clearly, a more sophisticated analysis is needed. One approach would be to add more bells and whistles to the simple analysis provided above (e.g., Lewis 2000); another would be to offer an analysis of a radically different sort (e.g., Mackie 1980). Here, however, there is no need to pursue these questions in detail. Let us simply say that a factor is linked to an outcome whenever the factor and the outcome stand in the relation — whatever it might be — that makes it correct to attribute the latter to the former. The word ‘linked’ serves us here as a placeholder that can be filled in, as needed, with a description of the right sort of relation.

Now, how are we to explain the fact that, e.g., people attribute the odor more to George’s failure to take out the trash than to Harry’s failure to take out the trash? The answer I want to propose is that both of these factors are linked to the outcome but that there is another important dimension along which they differ. This other dimension is the status of their contrast cases. The contrast case of George not taking out the trash is George taking out the trash; the contrast case of Harry not taking out the trash is Harry taking out the trash. But these two contrast cases seem quite different. The case of George taking out the trash seems highly relevant to the situation at hand. (After all, it is what George ought to have done.) But the case of Harry taking out the trash just seems bizarre. Why would he have done that? The only answer is that there was no reason for him to do it and that it’s hard to see why we should even be considering that possibility now. It is for this reason, I want to suggest, that we attribute the odor more to George’s omission than to Harry’s.

This point comes out especially clearly when we consider contrast cases that are truly outlandish. Consider, for example, the case in which a sudden nuclear attack destroys the whole city, along with the building in which George and Harry work. If such
a nuclear attack had occurred at just the right time, perhaps the bad odor would never have arisen. And yet, we certainly wouldn’t attribute the odor primarily to the absence of a sudden nuclear attack. The problem with factors like the absence of a sudden nuclear attack is that their contrast cases seem so strange, so totally absurd, as not to be relevant to attribution at all.

So it seems that there is something special about contrast cases like the case of George taking out the trash that makes them eligible for attributions. One might say that these cases are ‘more relevant to the situation,’ that they are ‘especially salient,’ that they seem ‘worth considering.’ Let us capture these imprecise ideas with an imprecise concept. We will say that certain contrast cases obtrude more than others. The hypothesis, then, will be that attribution isn’t simply a matter of finding factors that are linked to an outcome. The degree to which people attribute to a given factor is also determined in part by the degree to which its contrast case obtrudes. People attribute primarily to the linked factors whose contrast cases obtrude most, less to the factors whose contrast cases obtrude to a lesser degree and hardly at all to the factors whose contrast cases obtrude least.

A question now arises as to how people determine the degree to which a given case obtrudes. The answer seems to be that a number of different kinds of considerations go into this determination — some normative, others purely descriptive.

Let us begin with the purely descriptive considerations. One of these is the frequency with which a given type of event actually arises. Thus, if a particular type of event occurs every Saturday for two years and then suddenly, one Saturday, it does not occur, the case in which this type of event occurs on that Saturday as well will obtrude to a great degree. Related to this first consideration is the issue of expectation. Suppose that we all expect a given event to occur and that, even after it has clearly failed to occur, we still don’t know quite why it never happened. Then the case in which this event occurs will obtrude to a great degree. Presumably, there are other criteria like the ones I have listed here. In any case, I will make no attempt to enumerate them all.

Instead, I simply want to emphasize that normative considerations also play a role in determining the degree to which a given case obtrudes. Specifically, if a given type of case ought actually to arise, then the case in which it does arise will obtrude to a great
degree. So, for example, suppose we conclude that a person ought to have performed a behavior that he did not actually perform. We will find, then, that the contrast case in which he does perform that behavior will obtrude to a great degree. And, of course, this phenomenon is not limited to human behaviors. If we conclude that a computer ought to have printed out a particular document (when, in actuality, it did not), we will find that the contrast case in which the computer does print out the document obtrudes to a great degree.

For present purposes, the most important aspect of this view is that it takes certain normative considerations to be relevant to the question one is trying to answer when one makes an attribution. In making an attribution, one looks for a factor whose contrast case obtrudes. And one thing that can make a case obtrude is a normative consideration — namely, the degree to which it ought to have been actual. Thus, it turns out — at least on the view that we have been presenting — that one of things that makes it right to attribute a particular outcome to a particular factor is the normative status of the contrast case of that factor.

IV

Within social psychology, however, research has traditionally been dominated by a quite different framework. This framework says that an attribution is an answer to a purely descriptive question, indeed by a question that is more or less statistical in nature. The clearest expression of this framework is Harold Kelley’s (1967) famous slogan the person as scientist.

The phrase ‘person as scientist’ is most closely associated with the theory (discussed in more detail below) according to which attribution should be understood as in some way resembling the analysis of variance. But the phrase also captures a more general view about the nature of people’s attributions. This more general view says that, when people are wondering whether to attribute a particular outcome to a particular factor, they are asking the kind of question that can be answered using standard scientific methods (statistics, experimentation, etc.). The basic claim is that an attribution is something like a scientific hypothesis — not quite as rigorous or precise perhaps, but still best understood as a more informal version of the same basic idea.
Why would anyone adopt such a view? One of the most compelling reasons, it seems to me, is the striking success of the scientific methodologies in the various domains where they have been systematically applied. When we see how successful these methods have been at solving problems in the systematic sciences — physics, psychology, and so forth — we may be tempted to hypothesize that people use similar methods when they are trying to solve the kinds of problems they encounter in ordinary life. Here we seem to be relying on a tried-and-true strategy for generating hypotheses in cognitive science. When trying to determine how people solve a given problem, it often proves helpful to look for the optimal method of solving that problem (say, by trying to build a computer program that solves it optimally) and then to hypothesize that people actually use that method. Now, we do have good evidence that certain aspects of scientific methodology are, if not literally optimal, then certainly highly effective in solving a broad array of problems. Why should we not hypothesize, then, that people’s everyday attributions are generated by a process that resembles some aspect of scientific methodology?

The short answer is that people often use attribution to solve a very different kind of problem from the ones for which scientific methodologies were designed. Specifically, it seems clear that people’s attributions play a key role in the process by which they determine whether or not anyone is to blame for a given outcome. And when people are faced with this sort of problem — the problem of assessing blame — it makes perfect sense that normative considerations should play some role in the method by means of which they arrive at a solution.

Suppose, for example, that an accident has occurred and the government wants to know why. Presumably, the government would not simply be interested in developing a more refined scientific understanding of the phenomenon in question. It would be seeking an attribution because it needed to know whether or not anyone was to blame. So it will not be especially helpful for it merely to identify any old factor that is linked to the outcome. (It will not want to know, e.g., that the accident wouldn’t have occurred if the earth had been struck by an asteroid.) Instead, it will take a special interest in certain particular types of linked factors.
To the extent that our primary concern is with issues of blame, we will take a special interest in behaviors such that the agent did not do what he ought to have done. And, above all, we will be interested in contrasting such behaviors with the case in which the agent did do what he ought to have done. The importance of normative considerations in people’s attributions lies precisely in the fact that they help people to identify these specific types of linked factors — the factors that are most relevant to judgements of blame.

V

The key claim of the theory we have been developing thus far is that normative considerations actually play a role in the fundamental competence underlying people’s causal attributions. But, of course, we have no direct access to the workings of that competence. Our strategy has therefore been to start with people’s intuitions regarding specific cases and then to construct a theory that explains those intuitions.

The question now is whether it is possible to construct an alternative explanation — an explanation that accounts for all of the data about people’s intuitions without assigning any place to normative considerations in the competence underlying people’s attributions. We will consider three putative alternative explanations. In each case, the question will be whether the alternative explanation can do a better job of accounting for people’s intuitions than the theory we presented above.

Alternative Explanation 1: Conversational Pragmatics

Consider the position of a person who is being asked why a given outcome occurred. The person’s answer is unlikely to be simply a spontaneous reflection of his own attributions. Most likely, he will try to figure out what exactly his audience is after — what it is that they want to understand about the outcome in question — and he will construct his response in such a way that it provides the information they need to attain this understanding. This is the force of what I will call conversational pragmatics.

Now, in the kinds of cases that we have been discussing, it does seem quite simple for a person to construct a coherent story about why he is being asked for an attribution. So, for example, if an accident occurs and someone asks, “Whose actions
were the main cause of this accident?” it will usually be a good guess that the questioner really wants to know who is to blame for the accident. Perhaps, then, people’s answers don’t accurately reveal their actual attributions. Instead, their answers may reflect an attempt to respond to an implicit question about blame.

Continuing along this same basic line, one can easily construct an alternative explanation for the data reported above. Perhaps the apparent impact of normative considerations is, in fact, due entirely to conversational pragmatics. It might turn out that people’s attributions themselves are purely descriptive — with normative considerations entering the picture only where the audience is clearly interested in issues of blame. Then people’s responses to certain experimental questions would be influenced by normative considerations even if people’s attributions themselves were not normative in any way.

One way to determine which of the two explanations is correct would be to construct a case that did not have the usual sort of pragmatic significance — a case in which no one could reasonably suppose that the real point of the question was to determine who was to blame for a given outcome. Then we could check to see whether, even in a case of this latter type, people’s intuitions seemed to be influenced by normative considerations.

For example, suppose that we modify our story about the trash and the janitor in such a way that the outcome turns out to be purely positive:

George and Harry both work in a large office building. George is the janitor; Harry takes care of the mail.

Every day, George goes through the entire building and empties all of the garbage baskets. Since the building is large, this task normally takes him about one half hour.

One day, George is feeling tired and decides not to take out the garbage.

Harry sees that the garbage hasn’t been taken out. He doesn’t go to take it out himself, since that is not his job.

But it turns out that the company is extremely lucky. An accountant had accidentally thrown out an important document, and everyone is overjoyed to find that the trash hadn’t been taken out and hence that the document is still there.

Now suppose that we pose the question: “Why is the document still there?” Surely, this question would not normally be construed as a request for information about who is to
blame. (After all, the outcome was a good one, and it therefore seems that the question of blame does not even arise.) And yet, we would still be inclined to attribute the outcome primarily to George’s omission.

On the hypothesis provided above, this is exactly what one would expect. Since George ought to have taken out the trash, the contrast case in which George takes out the trash obtrudes, and people therefore attribute the outcome to his omission. Of course, the outcome in this case is a positive one, but there is no reason to suppose that the goodness or badness of the outcome will have any influence on people’s intuitions in a situation like this one. The claim is that cases in which the agent does what he ought to do will always obtrude, regardless of the nature of the outcome.

Perhaps there is some way to explain even this result in terms of conversational pragmatics, but I cannot think of what it might be. At this point, then, it seems that the best supported hypothesis would be the hypothesis according to which normative considerations truly do play a role in people’s attributions.

Alternative Explanation 2: Blame Validation

We turn now to Mark Alicke’s (2000) theory of blame validation. This theory states that people’s judgments of blame sometimes influence their causal attributions. More specifically, people’s feeling that an agent deserves blame for a given outcome sometimes leads them to attribute that outcome to the agent’s behavior. It can be seen immediately that this theory is similar to my own in at least one respect: both theories claim that people’s moral judgments can influence their causal attributions. But the similarity ends there. In most other respects, the two theories are radically different.

To understand the key differences, we need to make a clear distinction between the judgment that an agent’s behavior is wrong and the judgment that this agent is to blame for a given outcome. Often, the former judgment will require far less information than the latter. Thus, we might quickly conclude that it was wrong for Jack to give his 12 year-old sister cocaine, but it would be far more difficult to determine whether he is therefore to blame for the health problems she subsequently developed.
The theory of blame validation posits a process that works something like this:

`wrong` --- `blame` --- `attribution`

Here our intuitive sense that the behavior is wrong immediately leads us to blame the agent. The causal attribution then serves only to justify the blame we have already assigned.

I see no reason to doubt that this sort of process does sometimes take place. It certainly seems plausible that there might be a causal sequence whereby we first (a) feel horrified by Jack’s decision to give his little sister cocaine, then (b) immediately blame him for the health problems his sister later developed, and finally (c) attribute the health problems to his decision as a way of justifying the blame we have already assigned. Alicke and his colleagues have amassed an impressive array of evidence for the existence of processes like this one (e.g., Alicke 2000; Alicke, Davis, & Pezzo 1994).

Still, the existence of such processes gives us no reason to suppose that moral considerations are playing any role in the fundamental competencies underlying causal attribution. What we see here is simply a performance error. People’s fundamental competencies are being overridden and distorted by an attempt to justify pre-existing moral judgments.

Without denying that such distortions occasionally take place, I want to suggest that moral considerations also play a second, very different sort of role. In particular, they can be involved in causal sequences like this:

`wrong` --- `attribution` --- `blame`

Here our judgment that the behavior was wrong directly influences our causal attributions. These causal attributions then influence our judgments of blame.

The chief source of evidence for this type of process comes from cases in which people judge that the agent’s behavior is wrong but do not feel that the agent is to blame for a particular outcome. Take our case of the corporation that manages to hold on to an important document because George failed to take out the trash. If people attribute this
outcome to George’s behavior, it cannot be because they want to justify a judgment that George is to blame for the fact that the corporation still has the document. (There is no blame here to justify.) Rather, judgments about the wrongness of the behavior appear to be directly influencing people’s attributions — playing a role that has nothing to do with justifying a pre-existing judgment of blame.

To the extent that this sort of process appears to be taking place, we have at least prima facie evidence for a new view about the relationship between causal attributions and judgments of blame. It seems that people’s causal attributions are not simply serving to justify pre-existing blame judgments. Rather, causal attributions are sensitive to moral considerations in a way that makes it possible for people to use them in a later process of determining whether an agent is deserving of blame. This general picture opens up the possibility that moral considerations might really be playing a role in the fundamental competencies underlying people’s causal attributions.

Alternative Explanation 3: Analysis of Variance

There is a long and distinguished tradition in social psychology according to which attributions are best understood in terms of the statistical technique known as analysis of variance (ANOVA). This model was first proposed in a classic article by Kelley (1967); it has been extended in various ways by McArthur (1972), Pruitt and Insko (1980), Försterling (1989; 1992) and others, and it is now the standard textbook view.

To understand this view, it is perhaps best to begin by considering a case in which analysis of variance might actually used as a statistical technique and then to consider the ways in which this technique would have to be modified to form a plausible model of people’s attributions. First, the statistical technique. Suppose that we are wondering what factor is most important in determining the grades that students get on math tests. Are the differences between different students having a big impact, or is the most important factor the differences between different tests? To answer this question, we can consider a certain distribution of students and a certain distribution of tests. Then, for each possible pair of student and test, we need to know the grade the student received when given the test. It will then possible to determine in a precise and mathematical way which of the
two factors had more impact on the outcome — the student or the test. For example, suppose that the different students tend to get more or less the same grades when presented with the same test but that any given student gets radically different grades depending on which test he is given. Then it can be shown that the variance in grades is due primarily to differences between the tests, not to differences between students. In fact, it would be possible to compute exactly what percentage of the variance was due to differences between students and what percentage was due to the differences between tests.

Now consider a case in which we are trying to make an attribution for a particular outcome. Peter has just gotten a bad grade on a test, and we want to know whether to attribute this grade primarily to something about Peter or to something about the test of itself. Clearly, we would not normally solve such a problem by making any kind of rigorous mathematical calculation, but it is often suggested that people’s attributions do involve a more informal version of the statistical analysis of variance. The claim is that, when we are trying to make an attribution for Peter’s bad grade, we think about how Peter did on other tests and about how other students did on this test. If most other students did well on this test but Peter does poorly on most other tests, we attribute the grade primarily to Peter. Conversely, if Peter does well on most other tests and most other students did poorly on this test, we attribute his grade primarily to the test.

It is important to note that, on this sort of theory, attribution is not simply a matter of considering certain counterfactuals. Hence, our attribution for Peter’s bad grade is not supposed to be a matter of considering what would have happened if Peter’s test had been easier. On the contrary, the theory attaches a great deal of significance to questions about the actual distribution. In particular, the theory implies that our attribution for Peter’s grade will depend on how frequently it actually occurs that students receive difficult tests vs. how frequently they receive easy tests.

For a clearer example of this phenomenon, suppose that we were wondering whether to attribute Peter’s bad grade to the fact that the teacher was not struck by a lightning bolt on his way home from work. Perhaps it’s true that such a lightning bolt would have prevented the teacher from correcting Peter’s test (and hence prevented Peter from getting a bad grade), but under ordinary circumstances, we surely wouldn’t
conclude that the bad grade was due primarily to the absence of the lightning bolt. On the theory currently under discussion, this aspect of people’s attributions is to be understood in terms of the analysis of variance. The basic idea is that, when we consider the variance in students’ actual grades, we find that very little of this variance is explained by the presence or absence of lighting bolts — the reason being simply that deaths by lightning are too infrequent to have any substantial impact on the variance in students’ grades. Thus, the frequency with which a given type of factor arises can sometimes play a major role in people’s attributions (Försterling 1989). And in general, we are unlikely to be justified in attributing an outcome to a factor of some type whose contrast case almost never arises.

For our purposes, the most important point about this theory is that it describes people’s attributions as purely descriptive statistical judgements. Normative considerations play no role at all. And yet, it does seem that this framework can explain most of the intuitions that we have discussed thus far. So, for example, we earlier explained the difference between Harry and George in terms of people’s (normative) belief that Jake really ought to have come in to work. But we could just as easily have explained the effect in terms of beliefs about behavioral variance. The idea would be that, if someone else had been in Harry’s position, that other person would probably have done exactly what Harry did, whereas if someone else had been in George’s position, that person would probably have done something that prevented the bad odor from filling the building. Thus, George’s omission explains the outcome in a way that Harry’s does not.

The key point here is that, in the specific cases we have been discussing thus far, the agent who does something wrong is also doing something unusual, something that most people in his or her position would not have done. So to distinguish between the two explanations, we need to look at cases in which the property of being unusual somehow comes apart from the property of not being what the agent ought to have done.

What we really need, then, is a type of behavior that almost everyone performs but that almost everyone ought to stop performing. Unfortunately, it can be quite difficult to find a type of behavior that uncontroversially has this property. The best we can do, it seems, is to momentarily suspend all questions about which behaviors really ought to be performed and simply to accept, for the sake of argument, that a certain type of behavior
that almost everyone performs is morally wrong. In short, the goal is to enter into the
mindset of someone who believes that a certain type of behavior is morally wrong and
then to ask what attributions seem appropriate from the perspective of the view we have
temporarily adopted.

For a simple example, consider a vegan who believes that it is immoral to
consume meat, eggs, milk or any other animal product. Such a person presents us with an
ideal test case in the conflict between (a) the theory according to which attributions are
based entirely on an analysis of variance and (b) the theory according to which normative
considerations also play a role.

After all, it seems that adherents of theory (a) should reason as follows:

The fact that veganism is extremely rare will greatly limit the extent to which it can
figure as a factor in people’s attributions. It may be true of a given non-vegan that, if
he had been a vegan, he would thereby have brought about a different outcome from
the one that actually arose. But except in unusual cases, it would be absurd to attribute
the outcome that actually arose to the fact that this person was not a vegan. Since
almost everyone is not a vegan, the non-veganism of this one person is unlikely to
explain any substantial portion of the variance in the outcome variable.

By contrast, adherents of theory (b) can reason like this:

The vegan under discussion believes that everyone ought to be a vegan. When it
comes time to make an attribution, she will therefore find that contrast cases in which
other people become vegans obtrude to a great degree. So it will make perfect sense
for her to attribute certain outcomes to the fact that particular other people are not
vegans, even though she knows full well that almost everyone is not a vegan.

We can now ask which of these two chains of reasoning is best supported by the evidence
from people’s attributions.

Consider the case of a vegan whose best friend (a non-vegan) dies of cancer. Here
it seems quite plausible that she might say: “The main reason he got cancer is that he
never became a vegan.” Certainly, this sentence is not absurd in the way that, e.g., it
would be absurd to attribute Peter’s bad grade to the fact that his teacher wasn’t struck by
a lightning bolt. Quite the opposite: what would be truly absurd would be for us to
respond by saying, “But his cancer couldn’t possibly be due primarily to his not being a
vegan! After all, there are so few vegans…” Confronted with a response like this one, our
vegan could simply say: “Look, I know that almost everyone continues this immoral practice of consuming animal products. But I don’t see what that has to do with it. All I’m saying is that my friend’s death was due primarily to the fact that he didn’t switch over and become a vegan.”

Similar phenomena can be observed when people who are members of small cults attribute outcomes to the fact that one or another person is not a member of the cult. Or when people who hold unorthodox scientific opinions attribute outcomes to the fact that certain specific scientists do not hold those opinions. Or when people who are opposed to certain widely shared rules attribute outcomes to the existence of those rules within specific communities. These people are not somehow misunderstanding the nature of attribution. Rather, it seems that the nature of attribution is being misunderstood by those theorists who assume that normative considerations play no role and that all of these people must therefore be mistaken.

I conclude that the available evidence lends more support to the view that normative considerations play a role in attribution than to the view that attribution should be understood entirely in terms of the analysis of variance.

VI

It is widely believed, at least among social psychologists, that a causal attribution should be understood as something like a scientific hypothesis. Against this view, we have argued that causal attributions have an essential normative element — that they are concerned not only with what is the case but also with what ought to be the case.

Our arguments for this claim were based entirely on data about the attributions people make in particular cases. We did not try to argue from any more general theoretical position. Still, it seems that the issue we have been discussing here is related to a broader question about the relationship between judgments of blame and judgments about causal and psychological questions.

The results of our inquiry seem to suggest that some of our causal and psychological judgments are best understood as multi-purpose tools. That is, it appears that our capacity to make these judgments was shaped both by the need for ‘scientific’ understanding and by the need to arrive at ascriptions of blame. In other words, the
capacity works the way it does in part because, by working that way, it enables us to make blame ascriptions.

The evidence presented here seems to suggest that, for ordinary causal attributions, this ‘multi-purpose tool’ view might be correct. It will be interesting to see whether a similar picture can be applied to yet other aspects of our capacity to make causal and psychological judgments.


87


