Reason Explanation in Folk Psychology[1]

Joshua Knobe
University of North Carolina – Chapel Hill

(Draft of a paper to appear in *Midwest Studies in Philosophy*)

Consider the following explanation:

(1) George took his umbrella because it was just about to rain.

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, the explanation given above seems to be saying that George's reason for taking his umbrella was that it was just about to rain. 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.[2]

Thus far, research on these issues has been characterized by a very specific sort of methodology. 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 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 is being implemented by computer programs that were carefully engineered to do a perfect job of accomplishing some particular goal. The practice of reason explanation is something that has been constructed by *people*. There is therefore no guarantee that one will be able to derive every aspect of the practice from some grand theory about the nature of reasons and their role in explanation. Quite possibly, it will turn out that the only way to reach an adequate understanding of the practice is to think in more detail about how the human mind actually works.

My aim here is to pursue this alternative approach. I argue that there are features of the practice of reason explanation that are extremely difficult to accommodate on the traditional approach but that can be easily understood in light of certain facts about how people's minds work. In particular, it seems that people's intuitions about reason explanations can be affected by their *moral judgments*. That is to say, when people are evaluating a reason explanation, it seems that their intuitions can be affected by their judgments about whether certain behaviors are morally good or morally bad. This aspect of our practice may at first seem a bit bizarre, and it does not appear to follow from any plausible theory about the fundamental nature or purpose of reason explanation. Nonetheless, I think it is possible to show that the surprising patterns we find in people's intuitions actually follow in a fairly straightforward way from certain more general facts about human psychology.

I

One 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) a. Susan punished her son because he broke the vase.

b. Susan punished her son because she is an irritable person.

Intuitively, it seems that there is a fundamental difference between these two explanations. The first explanation is an attempt to give Susan's *reason* for punishing her son. By contrast, the second explanation is not an attempt to give Susan's own reasons; it simply describes a factor that caused her to behave as she did.

A second way to get a sense for the concept of reason explanation is to imagine a kind of inner monologue that the agent goes through before making a decision. An agent might think: 'He broke the vase; therefore I will punish him.' But she definitely would *not* think: 'I am an irritable person; therefore I will punish him.' Presumably, the thought that she is an irritable person does not enter into her practical reasoning in any way.

A third approach would be to make use of a metaphor. In particular, one wants to describe the phenomenon using either the metaphors of *ground* and *light*. Thus, one might say that Susan decided to punish her son 'on the grounds that he broke the vase' or 'in light of the fact that he broke the vase.' However, it seems wrong to say, e.g., that she punished him 'on the grounds that she is an irritable person.' That second type of explanation is not an attempt to give Susan's own grounds for acting; it works in an entirely different way.

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 the 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 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 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 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 stipulative 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 is being proposed 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 following sentences:

(3) a. He *splashed* the paint onto the wall.

b. He *smeared* the paint onto the wall.

It seems initially that these two sentences have more or less the same form, but the usual view among linguists is that they are actually very different.

This difference comes out clearly when we try to modify these sentences in certain ways. For example, suppose that we eliminate the agent from each of them (to form what is sometimes called the 'inchoative alternant'). We then get:

(4) a. The paint splashed onto the wall.

b. * The paint smeared onto the wall.

Here we find a difference: the first sentence remains acceptable; the second does not (Hale & Keyser 1993). The difference we see in (4) gives us at least some evidence that people were actually non-consciously classifying the verbs in (3) into two different categories.

The suggestion now is that something very similar is happening in people's use of different sorts of explanations. Thus, consider the explanations:

(5) a. He sent her that nasty letter *because he wanted to make her feel bad.*

b. He was disappointed to hear about her success *because he wanted to make her feel bad.*

These two explanations may initially seem to have the same form, but the claim is that people are non-consciously categorizing them in very different ways. Specifically, it is claimed that they are non-consciously determining that the first is a reason explanation and the second is not.

We can obtain some preliminary evidence for this conjecture simply by looking at how each of the explanations fare when we change some of the words around. For example, suppose that we eliminate from both explanations the words 'because he wanted.' Applying this modification to the first explanation, we get the perfectly acceptable:

(6) He sent her that nasty letter *to make her feel bad*.

But when we try to apply the same modification to the second explanation, we end up with the unacceptable:

(7) * He was disappointed to hear about her success *to make her feel bad.*

As we will see below, these facts about the actual syntactic forms of different kinds of explanations can help us gain some valuable insight into the nature of the explanations themselves.

But first we need to embed these preliminary comments within a richer theoretical framework. In particular, it will prove helpful to take certain ideas from what has come to be known as the *folk-conceptual theory* of behavioral explanations. (Knobe & Malle 2002; Malle, et al. 2000).

This theory claims that people ordinarily make sense of explanations by employing two distinct levels of representation – the level of 'linguistic structure' and the level of 'conceptual structure.' The linguistic structure of an explanation is the structure of words that are used to express it; the conceptual structure is the structure of concepts people use to interpret it. The essence of the folk-conceptual theory is that people use a complex system of rules to go back and forth between these two levels.

Thus, suppose a person hears the explanation:

(8) Because he wanted to make her feel bad.

The person could then use these rules to map this linguistic structure onto a conceptual structure that looked something like this:

Reason Explanation

Type: Desire

Content: To make her feel bad

But the mapping between linguistic structure and conceptual structure is not one-to-one. Hence, in a somewhat different context, that very same linguistic structure could be mapped onto a conceptual structure like:

Merely Causal Explanation

Causal Factor: He wanted to make her feel bad.

The upshot of all this is that many linguistic structures suffer from a kind of ambiguity. [3]

We can now introduce a claim that will play a crucial role in the argument that follows. It seems that there is a special class of linguistic structures that actually are not ambiguous in the usual way. The linguistic structures in this class do not get mapped onto different conceptual structures depending on the context. Regardless of the context, they always get mapped onto the very same conceptual structure. The linguistic

structures in this special class are usually referred to as *purposive clauses*. Here are a few examples:

(9) a. She did that *for John's sake*.

b. She did that *so that John would feel better*.

c. She did that *in order to make John feel better*.

The key point about these linguistic structures is that they can only be interpreted as reason explanations. The explanations given in (9) cannot be interpreted as asserting that the agent's desire to help John simply *caused* her to perform the behavior. They can only be interpreted as asserting that John actually figured in her *reason* for performing the behavior.

People's use of purposive clauses thereby offers us a window into the fundamental processes underlying everyday behavior explanations. Since purposive clauses can only be used to convey reason explanations, we can use people's intuitions about purposive clauses to get a better sense for the conditions under which they regard reason explanations as appropriate.

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.

Here it may be helpful to introduce a pair of vignettes that I originally constructed for a somewhat different purpose (Knobe 2003). First, consider a vignette about a corporate executive who harms the environment:

> 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.

And now suppose that we replace every instance of the word 'harm' with 'help,' yielding a vignette about an executive who helps the environment:

> 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.

In the present context, the important point about these two vignettes is that they elicit very different intuitions about the use of reason explanation (Knobe 2004). Faced with the first vignette, most people think it sounds right to use the reason explanation:

(10) The chairman of the board harmed the environment *in order to increase profits.*

But faced with the second vignette, most people do not think it sounds right to use the reason explanation:

(11) The chairman of the board helped the environment *in order to increase profits.*

Yet it seems that the only major difference between these two vignettes lies in the moral status of the agent's behavior.

In thinking about how to account for results like this one, researchers have arrived at a kind of limited convergence. Considerable disagreement remains about a number of issues (e.g., Adams & Steadman 2004; Knobe forthcoming; Nadelhoffer forthcoming; Turner 2004), but researchers working on these phenomena generally agree that moral considerations are somehow influencing people's intuitions about the proper use of reason explanations.

V

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 a natural sort of approach, which has been applied to a broad variety of philosophical problems.

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.

It is time for a radical change in perspective. Instead of thinking about the nature of reasons, we need to think about the nature of the human mind. Intuitions about reason explanation are a product of certain cognitive capacities. If we can understand those capacities, we may be able to understand why the practice of reason explanation works as it does.

The key thing to keep in mind here is that the relevant underlying capacities were not specifically engineered to do the best possible job at understanding reason explanations. In fact, many of these capacities are used not just for understanding reason explanations but in a whole variety of different tasks. Hence, it is possible to test hypotheses about these capacities using independent evidence – evidence that is not derived from intuitions about reason explanations.

More importantly, it is possible that the patterns we observe in people's intuitions about reason explanations will be susceptible to a kind of indirect explanation. When we simply look on the surface and observe the patterns themselves, they may seem utterly baffling and arbitrary. But suppose that we look instead at each of the underlying capacities. It may turn out that we can easily understand why each of these capacities

ended up working the way it does, and it may then happen that the interaction among these various capacities automatically generates (as a kind of by-product) certain seemingly-bizarre patterns in people's intuitions.

VI

Let us focus first on the capacity to distinguish between *intentional* and *unintentional* behavior. As numerous authors have noted, there seems to be an important connection between this capacity and the practice of reason explanation (Anscombe 1957; Goldman 1970; Malle et al. 2000; Mele 1992). Philosophers continue to debate 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 first thing to note here is that the asymmetry we observed in intuitions about reason explanations also emerges in intuitions about intentional action. Thus, most people think it sounds right to say:

(12) The chairman of the board *intentionally* harmed the environment.

But most people think it sounds wrong to say:

(13) The chairman of the board *intentionally* helped the environment.

So far, this is all precisely the same as the pattern we find in the case of reason explanation. But there is also an important difference. Unlike the asymmetry observed in intuitions about reason explanation, the asymmetry in intuitions about intentional action seems to make a certain amount of sense.

The basic idea here is a simple one. People often use the concept of intentional action to determine whether an agent deserves blame or praise for her actions. And when the asymmetry is seen in the context, one can immediately see why it might arise. It seems that the chairman deserves considerable blame in the vignette where he harms the environment but that he does not deserve any praise in the vignette where he helps the environment. There is ample reason to suppose that this fact about blame and praise lies at the heart of the asymmetry we find in intuitions about intentional action. Hence, although there are still a number of rival theories about precisely why the effect arises

(Adams & Steadman 2004; Alicke forthcoming; Knobe forthcoming; Nadelhoffer forthcoming a; Nichols & Ulatowski 2006), almost all researchers now agree that it has something to do with the difference between the conditions under which people assign blame and those in which assign praise.

VII

What we have seen thus far is a certain kind of parallel between intuitions about reason explanation and intuitions about intentional action. The key question now is how this parallel is to be explained.

The traditional approach to this question was to suppose that we have some independent notion of reason explanation and that we can then use that notion to construct an account of intentional action. So, for example, Anscombe (1957) suggests that there is a particular type of question that is best understood as a request for a reason explanation. She then defines the word 'intentional' by saying that an intentional action is a behavior about which this type of question can be legitimately asked.

It seems to me that this sort of approach cannot adequately account for the data we have amassed thus far. We have seen that there is an asymmetry in intuitions about intentional action and a parallel asymmetry in intuitions about reason explanation, but we have also seen that the two asymmetries differ in a crucial respect. The difference is this: there are a variety of plausible hypotheses about how moral judgments might be directly influencing intuitions about intentional action, but no one has been able to come up with any plausible hypothesis about how moral judgments might be directly influencing intuitions about reason explanation.

In light of these considerations, it seems more than a little perverse to suggest that we have some prior understanding of the notion of reason explanation and that we can then just define the concept of intentional action in terms of this prior understanding. Surely, it would be more plausible to suggest that the process goes in the opposite direction. People start out with a concept of intentional action (a concept in which moral considerations play a key role) and then use that concept in the psychological process underlying intuitions about reason explanation.

Perhaps the basic idea behind this hypothesis is best conveyed by means of a fable.

> 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.'
>
> This second person eventually prevails. Since moral considerations already played a role in the concept of intentional action, this decision make it the case that moral considerations come to play a role in reason explanation as well.

Clearly, many aspects of this fable are purely figurative, but the claims about relationships between concepts are meant to be taken in the most literal way possible. Thus, the view is not simply that there is a certain sort of correlation between intuitions about intentional action and intuitions about reason explanation. Rather, the view is that there is literally a non-conscious process whereby moral judgments influence intuitions about intentional action which then serve as input to the mechanisms underlying intuitions about reason explanation.

At this point, however, we need to consider a possible objection. The objection is that people's intuitions about reason explanations do not perfectly track their intuitions about intentional action. In particular, there appear to be cases in which people are willing to apply reason explanations to a behavior even though they do not regard that behavior as intentional.

This problem comes out especially clearly when we consider cases in which the agent himself believes that he is bringing about a bad side-effect but most people would think that the side-effect in question was actually morally good. 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.

Faced with this vignette, most subjects say that the terrorist *did not* save the Americans 'intentionally' but that he *did* save the Americans 'in order to rescue his son' (Knobe & Kelly 2006). In other words, it simply isn't true that people only accept reason explanations for behaviors that they regard as intentional.

Results like this one leave us with a difficult theoretical predicament. On one hand, it seems that intuitions about reason explanation *almost always* track intuitions about intentional action; on the other, it seems that intuitions about reason explanation do not *always* track intuitions about intentional action – there are cases in which the two kinds of intuitions diverge. What we need now is a theory that explains why the two

types of intuition almost always go together but also explains how there can be exotic cases in which they diverge.

When the divergences were first discovered, it did not seem possible to provide a theory that met this requirement. However, recent research has led to the construction of a new and very different theoretical framework. Within this new framework, it is possible to provide a simple and unified account that predicts both the intuitions about intentional action intuitions and the intuitions about reason explanation. The basic idea will be that there is a single sort of mechanism underlying both of these intuitions but that this mechanism yields different results depending on the nature of the question one is considering.

IX

But before we introduce this new framework, we need to revisit the vexed question of precisely how it is that moral considerations influence intuitions about intentional action. My original hypothesis was that people actually make a *judgment that the side-effect is morally bad* and that this judgment then influences their intuitions about whether the agent acts intentionally. Within the context of this basic framework, it was extremely difficult to come up with any plausible explanation for people's intuitions in the terrorist case. Regardless of whether people are asked the question about whether he did it 'intentionally' or the question about whether he did it 'in order to...,' they should conclude that there was nothing bad about saving the Americans. So it seems that this hypothesis leads immediately to the prediction that people's intuitions about reason explanation will always track their intuitions about intentional action.

Thankfully, more recent research has led to the creation of a radically different theoretical framework in which the experimental results can be more readily explained. First of all, a series of experimental studies by a variety of authors have conclusively refuted my original hypothesis (Cushman 2006; Machery 2006; Nichols unpublished data; Phelan & Sarkissian 2006; Pizarro, et al. unpublished data; Sinnott-Armstrong, et al. 2006; Wright & Bengson 2006). These studies have demonstrated to everyone's satisfaction that the impact of moral considerations on intuitions about intentional action cannot simply be understood in terms of people's judgments about certain side-effects

being morally bad. But the studies have also yielded an even more surprising result. Taken collectively, they seem to indicate that the overall pattern of people's intentional action intuitions does not track *any* of the moral judgments we are consciously aware of making. In other words, people's intuitions about intentional action do appear to be influenced in some way by moral considerations, but it doesn't seem possible to pick out any particular sort of moral judgment that we are aware of making and say: 'It is *this* sort of moral judgment that is influencing people's intuitions about intentional action.'

The secret to untangling this confusion was first suggested by David Pizarro and Paul Bloom. They pointed out that it might be possible to explain all of the data if we posit certain *non-conscious moral judgments*. The basic idea behind this framework is a simple one. When people observe a behavior, they have an immediate, intuitive reaction that leads to a non-conscious judgment. Then, when they have time to think about it in more detail, they can reflect on a variety of additional considerations and gradually form a considered conscious judgment. This conscious judgment may directly contradict the non-conscious judgment that they reached originally, but the non-conscious judgment is not thereby deleted from memory. It remains in existence and can continue to influence certain further psychological processes. In particular, it can influence people's intuitions about intentional action.

Together, Pizarro, Bloom and I conducted a study to test this hypothesis. To start out with, we needed to select a domain in which people's non-conscious moral judgments might diverge from the conscious ones. We chose judgments about gay kissing and interracial sex. Although many people consciously believe that there is nothing at all wrong with these behaviors, we speculated that people might be making non-conscious judgments that these behaviors are actually morally wrong. In the study, subjects were therefore randomly assigned either to receive vignettes about agents who encourage gay kissing and interracial sex or to receive vignettes about agents who encourage completely innocuous behaviors. (So, for example, some subjects were asked about gay men french-kissing on the street while others were asked about 'couples' french-kissing on the street.) As expected, subjects who received the vignettes about gay kissing or interracial sex were more inclined to say that the agent acted intentionally. Moreover, there was a significant correlation such that subjects who were high in a dispositional tendency to experience

disgust were especially likely to regard these behaviors as intentional (Pizarro, et al. 2006). These data offer some preliminary support for the view that the effect is actually driven by non-conscious judgments.

Building on this earlier study, we can now introduce a more detailed hypothesis about precisely how non-conscious moral judgments differ from conscious ones. The basic idea will be that non-conscious judgments are formed extremely quickly and therefore involve very shallow processing. In reaching a conscious moral judgment, we can consider a variety of different moral norms, weigh these norms against each other, perhaps even determine that some of the norms are themselves unjustified. Non-conscious moral judgments are formed through a far simpler process. In generating a non-conscious moral judgment, the only norms we consider are the ones that first come to mind. We do not search for additional norms; we do not weigh norms against each other; we do not ask whether any of the norms might themselves be unjustified. Instead, we simply determine whether the behavior in question violates any of the norms in the very limited set we are considering. If it does, we classify it as a *transgression*.[4] It is this judgment as to whether or not the behavior is a transgression that then influences our intuitions about intentional action.

Perhaps it would be helpful here to consider a single case in more detail and explore step by step how the postulated processes might operate. Suppose that we are thinking about a society governed by some morally abhorrent law (say, a law according to which one is obliged to kill people of certain races). Now consider what might happen if we learned that a member of this society violated the law. As soon as we encountered this case, we would begin a rapid non-conscious process of evaluation that only made use of the most salient norms. If the law itself was made salient in the context, the law would be the most salient norm and the behavior would therefore be classified as a transgression. Subsequently, we might take a moment to reflect and consciously think about whether the agent's behavior was right or wrong. In that subsequent process, we would determine that the law itself was morally abhorrent and that there was nothing wrong with violating it. But this subsequent reflection would not alter our initial non-conscious judgment. That judgment would remain in place and would continue to influence our intentional action intuitions.

In the Appendix, I report results from a new experiment that appear to lend support to this basic picture. All subjects were given a story about a society governed by a 'racial identification law,' which serves to identify people of certain races so that they can be rounded up and sent to concentration camps. Some subjects were told that a particular corporate executive *violated* the requirements of this law; others were told that the executive *fulfilled* the requirements of the law. In both cases, subjects were told that the executive did not care at all about the law and was just trying to make as much money as possible. The key question was whether subjects would say that the executive acted intentionally. The results showed a dramatic asymmetry. Subjects tended to think that the executive violated the requirements intentionally but fulfilled them unintentionally.

This result is not explained by theories that rely entirely on conscious judgments. These theories typically say that subjects will regard side-effects as intentional when they consciously believe them to be morally reprehensible. Yet, what we have here is a case in which subjects consciously believe that violating the requirements is actually a good thing and nonetheless end up concluding that the agent acted intentionally. The effect here is therefore better explained on a theory that posits non-conscious judgments about transgressions.

It should be emphasized that I do not take the theory of transgression detection to be relevant only to questions about intentional action intuitions. Presumably, what we are uncovering here is a general truth about how people make moral judgments. The significance of intentional action intuitions is just that they are influenced by non-conscious transgression judgments and therefore provide valuable evidence about how those judgments are made.

X

We can now return to the case of the terrorist and the Americans. What one finds in that case is a conflict between two norms. On one hand, there are the norms that the subjects themselves hold; on the other, there are the norms that the subjects attribute to the agent. The agent then performs a behavior that is morally good according to one of these norms but morally bad according to the other. The question now is whether this behavior will be classified as a transgression.

The answer seems to be that the classification of the behavior will depend on which norm is most salient. If the perceived norm of the agent is most salient, the behavior will likely be classified as a transgression. If the subjects' own norm is most salient, it probably won't. Thus, anything that affects the salience of these different norms will affect the likelihood that the behavior is classified as a transgression.

Here, at last, one sees a way of explaining the observed divergence between intuitions about intentional action and intuitions about reason explanation. The key idea will be that the sentences people are being asked to evaluate can themselves affect the salience of certain norms. In this case, the relevant sentences were:

(14) The terrorist saved the Americans in order to rescue his son.

(15) The terrorist saved the Americans intentionally.

A sentence like (14) would typically be used in a conversation aimed at understanding the terrorist's own mental states and therefore makes salient the perceived norms of the terrorist. By contrast, a sentence like (15) would typically be used in a conversation about whether the terrorist deserves praise or blame and therefore makes salient the subjects' own norms. Hence, the very same behavior gets classified as a transgression in the context of one of these sentences but not of the other. This divergence in transgression judgments then accounts for the observed divergence in intuitions.

If this explanation turns out to be on the right track, we can account for the divergence between intuitions about reason explanation and intuitions about intentional action without introducing any additional ad hoc assumptions. The pattern of divergence would follow in a principled way from a general theory about the nature of transgression detection.

XI

Here we return to the peculiar character of the present account. In the discussion thus far, we have not been presenting anything that might be called a 'theory of reasons' or a 'theory of intentional action.' Instead, we have been constructing theories about certain underlying cognitive capacities – a theory about the mapping between linguistic structure and conceptual structure, a theory about the process of transgression detection,

and so on. Each of these capacities is treated as a distinct aspect of the human mind. The patterns observed in people's intuitions are then derived from general truths about the nature of the capacities themselves and the ways in which they interact.

Some may feel that, in presenting theories of this type, I am turning away from precisely those issues that are most important. They may say that the truly deep questions here are the metaphysical questions about the nature and status of reasons themselves and that I am simply being detained by various unimportant psychological details.

It seems to me that this perspective suffers from an overly constrictive conception of philosophical depth. Recent work in experimental philosophy may not be moving *outward* to an understanding of certain objects and properties in the world, but it is moving *inward* to an ever greater understanding of the fundamental processes underlying people's intuitive judgments. The results thereby offer us a deeper understanding of human nature – and that, I would argue, is justification enough.

Appendix

This appendix describes the experiment that was briefly mentioned in section IX.

Subjects were 41 people spending time in a Manhattan public park. Each subject was randomly assigned either to the *violate condition* or to the *fulfill condition*. Subjects in the violate condition received the following vignette:

> In Nazi Germany, there was a law called the 'racial identification law.' The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps.
>
> Shortly after this law was passed, the CEO of a small corporation decided to make certain organizational changes.
>
> The Vice-President of the corporation said: "By making those changes, you'll definitely be increasing our profits. But you'll also be violating the requirements of the racial identification law."
>
> The CEO said: "Look, I know that I'll be violating the requirements of the law, but I don't care one bit about that. All I care about is making as much profit as I can. Let's make those organizational changes!"
>
> As soon as the CEO gave this order, the corporation began making the organizational changes.

Subjects in the fulfill condition received a vignette that was almost exactly the same, except that the CEO fulfills the requirements instead of violating them:

> In Nazi Germany, there was a law called the 'racial identification law.' The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps.
>
> Shortly after this law was passed, the CEO of a small corporation decided to make certain organizational changes.
>
> The Vice-President of the corporation said: "By making those changes, you'll definitely be increasing our profits. But you'll also be fulfilling the requirements of the racial identification law."
>
> The CEO said: "Look, I know that I'll be fulfilling the requirements of the law, but I don't care one bit about that. All I care about is making as much profit as I can. Let's make those organizational changes!"

> As soon as the CEO gave this order, the corporation began making the organizational changes.

After reading these vignettes, subjects were asked three questions:

> (1) Did the CEO *intentionally* violate [fulfill] the requirements of the law?
>
> (2) How much blame of praise does the CEO deserve for what he did?
>
> (3) How much blame or praise would the CEO have deserved if he had been specifically trying to violate [fulfill] the requirements of the racial identification law?

For the first question, subjects were given two options marked 'yes' and 'no.' For the second and third questions, subjects marked their answers on a scale from -3 ('a lot of blame') to +3 ('a lot of praise'), with the 0 point marked 'no blame or praise.'

For the question about how much blame or praise the CEO deserved, there was no significant difference between responses in the violate condition (M = -.9) and the fulfill condition (M = -1.7). For the question about how much blame or praise the CEO would have deserved if he had been specifically trying, responses in the violate condition (M = .3) were significantly higher than those in the fulfill condition (M = -1.8), $t(39) = 3.2$, $p < .05$.

The key question was whether subjects would say that the agent acted intentionally in each of the two conditions. There, 81% of subjects in the violate condition said that he violated the requirements intentionally, whereas only 30% of subjects in the fulfill condition said that he fulfilled the requirements intentionally. This difference is statistically significant, $\chi^2(1, N = 41) = 10.8$, $p = .001$.

Judgments as to whether the agent acted intentionally were not significantly correlated with either of the explicit moral judgments (both ps >.7).

Notes

[1] For helpful comments on this paper, I am grateful to Fiery Cushman, Edouard Machery and G. F. Schueler. Then, on a more general level, I am grateful to endless hours of conversation with Shaun Nichols both for a number of important substantive suggestions and for the contagious sense of intellectual excitement that he brings to everything in experimental philosophy.

[2] 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).

[3] At this point, one might well ask what evidence we have for a distinct level of conceptual structure. The answer is that the notion of conceptual structure is a crucial part of a theory that, taken as a whole, can predict and explain a variety of important phenomena. 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 et al. 2000; O'Laughlin & Malle 2002).

[4] The name is due to Shaun Nichols (personal communication).

References

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

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

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

Anscombe, G. E. M. 1957. *Intention*. Second Edition. Ithaca: Cornell University Press.

Cushman, F. 2006. "Judgments of Morality, Causation and Intention: Assessing the Connections." Unpublished manuscript. Harvard University.

Goldman, A. 1970. *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall.

Hale, K. & Keyser, J. 1993. "On Argument Structure and the Lexical Representation of Syntactic Relations." In K. Hale and J. Keyser, editors, *The View from Building 20*, pages 53-110. MIT Press.

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

Knobe, J. & Kelly, S. 2006. "Can One Act for a Reason without Acting Intentionally?" Unpublished manuscript. UNC-Chapel Hill.

Knobe, J. & Malle, B. F. 2002. "Self and Other in the Explanation of Behavior." *Psychologica Belgica* 42: 113-130.

Malle, B. F., Knobe, J., O'Laughlin, M., Pearce, G., & Nelson, S. 2000. "Conceptual Structure and Social Functions of Behavior Explanations." *Journal of Personality and Social Psychology* 79: 309-326.

Mele, A. 1992. "Acting for Reasons and Acting Intentionally." *Pacific Philosophical Quarterly* 73: 355-74.

Nadelhoffer, T. forthcoming a. "Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Jury Impartiality." *Philosophical Explorations*.

Nadelhoffer, T. forthcoming b. "On Saving the Simple View." *Mind and Language.*

Nichols, S. & Ulatowski, J. 2006. "Intuitions and Individual Differences: The Knobe Effect Revisited." Unpublished manuscript. University of Utah.

Pizarro, D., Uhlmann, E., Tannenbaum, D. & Ditto, P. Unpublished Data. Cornell University.