Person as Scientist, Person as Moralist[1]

Joshua Knobe
Yale University

]

It has often been suggested that people's ordinary capacities for folk psychology and causal cognition make use of much the same methods one might find in a formal scientific investigation. A series of recent experimental results offer a challenge to this widely-held view, suggesting that people's *moral* judgments can influence the intuitions they hold both in folk psychology and in moral cognition. The present target article argues that these effects are best explained on a model according to which moral considerations actually figure in the fundamental competencies people use to make sense of the world.

Consider the way research is conducted in a typical modern university. There are departments for theology, drama, philosophy… and then there are departments specifically devoted to the practice of *science*. Faculty members in these science departments generally have quite specific responsibilities. They are not supposed to make use of all the various methods and approaches one finds in other parts of the university. They are supposed to focus on observation, experimentation, the construction of explanatory theories.

Now consider the way the human mind ordinarily makes sense of the world. One plausible view would be that the human mind works something like a modern university. There are psychological processes devoted to religion (the mind's theology department), to aesthetics (the mind's art department), to morality (the mind's philosophy department) … and then there are processes specifically devoted to questions that have a roughly 'scientific' character. These processes work quite differently from the ones we use in thinking about, say, moral or aesthetic questions. They proceed using more or less the same sorts of methods we find in university science departments.

This metaphor is a powerful one, and it has shaped research programs in many different areas of cognitive science. Take the study of *folk psychology*. Ordinary people have a capacity to ascribe mental states (beliefs, desires, etc.), and researchers have sometimes suggested that people acquire this capacity in much the same way that scientists develop theoretical frameworks (e.g., Gopnik & Wellman 1992). Or take *causal cognition*. Ordinary people have an ability to determine whether one event caused another, and it has been suggested that they do so by looking at the same sorts of statistical information scientists normally consult (e.g., Kelley 1967). Numerous other fields have taken a similar path. In each case, the basic strategy is to look at the methods used by professional research scientists and then to hypothesize that people actually use similar methods in their ordinary understanding. This strategy has clearly led to many important advances.

Yet, in recent years, a series of experimental results have begun pointing in a rather different direction. These results indicate that people's ordinary understanding does not proceed using the same methods one finds in the sciences. Instead, it appears that people's intuitions in both folk psychology and causal cognition can be affected by *moral* judgments. That is, people's judgments about whether a given action truly is morally good or bad can actually affect their intuitions about what that action caused and what mental states the agent had.

These results come as something of a surprise. They do not appear to fit comfortably with the view that certain aspects of people's ordinary understanding work much like a scientific investigation, and a question therefore arises about how best to understand them.

One approach would be to suggest that people truly are engaged in an effort to pursue something like a scientific investigation but that they simply aren't doing a very good job of it. Perhaps the competencies underlying people's judgments actually are purely scientific in nature, but there are then various additional factors that get in the way of people's ability to apply these competencies correctly. Such a view might allow us to explain the patterns observed in people's intuitions while still holding onto the basic idea that people's capacities for thinking about psychology, causation, etc. can be understood on the model of a scientific investigation.

This approach has a strong intuitive appeal, and recent theoretical work has led to the development of specific hypotheses that spell it out with impressive clarity and precision. There is just one problem. The actual experimental results never seem to support these hypotheses. Indeed, the results point toward a far more radical view. They suggest that moral considerations actually figure in the competencies people use to make sense of human beings and their actions.

1. Introducing the Person-as-Scientist Theory

In the existing literature on causal cognition and theory-of-mind, it has often been suggested that people's ordinary way of making sense of the world is in certain respects analogous to a scientific theory (Churchland 1981; Gopnik & Meltzoff 1997; Sloman 2005). This is an important and provocative suggestion, but if we are to grapple with it properly, we need to get a better understanding of precisely what it means and how experimental evidence might bear on it.

1.1. Ordinary understanding and scientific theory

To begin with, we will need to distinguish two different aspects of the claim that people's ordinary understanding is analogous to a scientific theory. First, there is the claim that human thought might sometimes take the form of a *theory*. To assess this first claim, one would have to pick out the characteristics that distinguish theories from other sorts of knowledge structures and then ask whether these characteristics can be found in ordinary cognition. This is certainly a worthwhile endeavor, but it has already been pursued in a considerable body of recent research (e.g., Carey & Spelke 1996; Goldman 2006; Murphy & Medin 1985), and I will have nothing further to say about it here. Instead, the focus of this target article will be on a second claim, namely, the claim that certain facets of human cognition are properly understood as *scientific*.

To begin with, it should be emphasized that this second claim is distinct from the first. If one looks to the usual sorts of criteria for characterizing a particular knowledge structure as a 'theory' (e.g., Premack & Woodruff 1978), one sees immediately that these criteria could easily be satisfied by, for example, a religious doctrine. A religious doctrine could offer systematic principles; it could posit unobservable entities and processes; it could yield definite predictions. For all these reasons, it seems perfectly reasonable to say that a religious doctrine could give us a certain kind of 'theory' about how the world works. Yet, although the doctrine might offer us a theory, it does not appear to offer us a specifically *scientific* theory. In particular, it seems that religious thinking often involves attending to different sorts of considerations from the ones we would expect to find in a properly scientific investigation. Our task here, then, is to figure out whether certain aspects of human cognition qualify as 'scientific' in this distinctive sense.

One common view is that certain aspects of human cognition do indeed make use of the very same sorts of considerations we find in the systematic sciences. So, for example, in work on causal cognition, researchers sometimes proceed by looking to the statistical methods that appear in systematic scientific research and then suggesting that those same methods are at work in people's ordinary causal judgments (Gopnik et al. 2004; Kelley 1967; Woodward 2004). Different theories of this type appeal to quite different statistical methods, but these differences will not be relevant here. The thing to focus on is just the general idea that people's ordinary causal cognition is in some way analogous to a scientific inquiry.

And it is not only the study of causal cognition that proceeds in this way. A similar viewpoint can be found in the theory-of-mind literature (Gopnik & Meltzoff 1997), where it sometimes goes under the slogan 'Child as Scientist.' There, a central claim is that children refine their understanding of the mind in much the same way that scientists refine their theories. Hence, it is suggested that we can look at the way Kepler developed his theory of the orbits of the planets and then suggest that children use the same basic approach as they are acquiring the concept of belief (Gopnik & Wellman 1992). Once again, the idea is that the cognitive processes people use in ordinary life show a deep similarity to the ones at work in systematic science.

It is this idea that we will be taking up here. Genuinely scientific inquiry seems to be sensitive to a quite specific range of considerations and to take those considerations into account in a highly distinctive manner. What we want to know is whether certain aspects of ordinary cognition work in more or less this same way.

1.2. Refining the question

But now it might seem that the answer is obvious. For it has been known for decades that people's ordinary intuitions show certain patterns that one would never expect to find in a systematic scientific investigation. People make wildly inappropriate inferences from contingency tables, show shocking failures to properly detect correlations, display a tendency to attribute causation to whichever factor is most perceptually salient (Chapman & Chapman 1967; McArthur & Post 1977; Smedslund 1963). How could one possibly reconcile these facts about people's ordinary intuitions with a theory according to which people's ordinary cognition is based on something like a scientific methodology?

The answer, I think, is that we need to interpret that theory in a somewhat more nuanced fashion. The theory is not plausibly understood as an attempt to describe all of the factors that can influence people's intuitions. Instead, it is best understood as an attempt to capture the 'fundamental' or 'underlying' nature of certain cognitive capacities. There might then be various factors that interfere with our ability to apply those capacities correctly, but the existence of these additional factors would in no way impugn the theory itself.

To get a rough sense for the strategy here, it might be helpful to return to the comparison with religion. Faced with a discussion over religious doctrine, we might say: 'This discussion isn't best understood as a kind of scientific inquiry; it is something else entirely. So if we find that the participants in this discussion are diverging from proper scientific methods, the best interpretation is that they simply weren't trying to use those methods in the first place.' This would certainly be a reasonable approach to the study of religious discourse, but the key claim of the person-as-scientist approach is that it would

not be the right approach to understanding certain aspects of our ordinary cognition. Looking at these aspects of ordinary cognition, a defender of the person-as-scientist view would adopt a very different stance. For example, she might say: 'Yes, it's true that people sometimes diverge from proper scientific methods, but that is *not* because they are engaging in some fundamentally different sort of activity. Rather, their underlying capacities for causal cognition and theory-of-mind really are governed by scientific methods; it's just that there are also various additional factors that get in the way and sometimes lead people into errors.'

Of course, it can be difficult to make sense of this talk of certain capacities being 'underlying' or 'fundamental,' and different researchers might unpack these notions in different ways:

- One view would be that people have a *domain-specific capacity* for making certain kinds of judgments but then various other factors intrude and allow these judgments to be affected by irrelevant considerations.
- Another would be that people have a *representation of the criteria* governing certain concepts but that they are not always able to apply these representations correctly.
- A third would be that the claim is best understood *counterfactually*, as a hypothesis about how people would respond if they only had sufficient cognitive resources and freedom from certain kinds of biases.

I will not be concerned here with the differences between these different specific views. Instead, let us introduce a vocabulary that allows us to abstract away from these details and talk about this approach more generally. Regardless of the specifics, I will say that

the approach is to posit an underlying *competence* and then to posit various additional factors that get in the way of people's ability to apply that competence correctly.

With this framework in place, we can now return to our investigation of the impact of moral considerations on people's intuitions. How is this impact to be explained? One approach would be to start out by finding some way to distinguish people's underlying competencies from the various interfering factors. Then one could say that the competencies themselves are entirely scientific in nature but that the interfering factors then prevent people from applying these competencies correctly and allow moral considerations to affect their intuitions. This strategy is certainly a promising one, and we will be discussing it in further detail below. But it is important to keep in mind that we also have open another, very different option. It could always turn out that there simply is no underlying level at which the relevant cognitive capacities are purely scientific, that the whole process is suffused through and through with moral considerations.

2. Intuitions and moral judgments

Before we think any further about these two types of explanations, we will need to get a better grasp of the phenomena to be explained. Let us begin, then, just by considering a few cases in which moral considerations appear to be impacting people's intuitions.

2.1. Intentional action

Perhaps the most highly studied of these effects is the impact of people's moral judgments on their use of the concept of *intentional action*. This is the concept people use to distinguish between behaviors that are performed intentionally (e.g., hammering in a nail) and those that are performed unintentionally (e.g., accidentally bringing the hammer down on one's own thumb). It might at first appear that people's use of this distinction depends entirely on certain purely scientific facts about the role of the agent's mental states in his or her behavior, but experimental studies consistently indicate that something more complex is actually at work here. It seems that people's moral judgments can somehow influence their intuitions about whether a behavior is intentional or unintentional.

To demonstrate the existence of this effect, we can construct pairs of cases that are exactly the same in almost every respect but differ in their moral status.[2] For a simple example, consider the following vignette:

> 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.

Faced with this vignette, most subjects say that the chairman *intentionally* harmed the environment. One might initially suppose that this intuition relies only on certain facts about the chairman's own mental states, e.g., the fact that he specifically knew his

behavior would result in environmental harm. But the data suggest that something more is going on here. For people's intuitions change radically when one alters the moral status of the chairman's behavior by simply replacing the word 'harm' with 'help':

> 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.

Faced with this second version of the story, most subjects actually say that the chairman *unintentionally* helped the environment. Yet it seems that the only major difference between the two vignettes lies in the moral status of the chairman's behavior. So it appears that people's moral judgments are somehow impacting their intuitions about intentional action.

Of course, it would be unwise to draw any strong conclusions from the results of just one experiment, but this basic effect has been replicated and extended in numerous further studies. To begin with, subsequent experiments have further explored the harm and help cases to see what exactly about them leads to the difference in people's intuitions. These experiments suggest that that moral judgments truly are playing a key role, since participants who start out with different moral judgments about the act of harming the environment end up arriving at different intuitions about whether the chairman acted intentionally (Tannenbaum, et al. 2009). But the effect is not limited to

vignettes involving environmental harm: it emerges when researchers use different cases (Cushman & Mele 2008; Knobe 2003a) and even when they turn to cases with quite different structures that do not involve side-effects in any way (Knobe 2003b; Nadelhoffer 2005). Nor does the effect appear to be limited to any one particular population: it emerges when the whole study is translated into Hindi and conducted on Hindi-speakers (Knobe & Burra 2006) and even when it is simplified and given to four-year old children (Leslie, Knobe & Cohen 2006). At this point, there is really a great deal of evidence for the claim that people's moral judgments are somehow impacting their intuitions about intentional action.

Still, as long as all of the studies are concerned only with intuitions about intentional action specifically, it seems that our argument will suffer from a fatal weakness. For someone might say: 'Surely, we have very strong reason to suppose that the concept of intentional action works in more or less the same way as the other concepts people normally use to understand human action. But we have good theories of many of these other concepts – the concepts of deciding, wanting, causing, and so forth – and these other theories do not assign any role to moral considerations. So the best bet is that moral considerations do not play any role in the concept of intentional action either.' In my view, this is actually quite a powerful argument. Even if we have strong evidence for a certain view about the concept of intentional action specifically, it might well make sense to abandon this view in light of theories we hold about various other, seemingly similar concepts.

In a way, the argument under discussion here is reminiscent of the strategy that American troops adopted during the Vietnam War. In the early stages of the war, the

Vietcong would try launching attacks on individual American bases, but the Americans were generally able to fend them off. After all, the Americans might sometimes be outnumbered at one particular base, but they had a large number of different bases, and when things got rough, they could always call on nearby bases for reinforcements. This strategy initially proved highly effective. But, of course, the Americans did not end up winning the war. The turning point came with the famous Tet Offensive, when the Vietcong launched a surprise attack on all of the American bases at the same time. Then none of the bases could bring in reinforcements from any of the others, and the progress of the war changed irreparably.

In just the same way, it seems that we will never be able to dislodge the prevailing view of the mind if we simply launch piecemeal attacks on theories of particular individual concepts. If we attack the prevailing view about the concept of intentional action, someone can always just say: 'But that approach worked so well when we applied it to the concept of causation!' And, conversely, when we attack the prevailing view about the concept of causation, someone can always say: 'But that approach worked so well when we applied it to the concept of intentional action!' The only way to make progress here is to launch a kind of theoretical Tet Offensive in which we provide evidence against a large swath of such theories all at the same time. Then no theory can be brought in as back-up because they will all be simultaneously under attack.

2.2. Further psychological states

To begin with, we can show that the effect observed for intuitions about intentional action does not arise only for people's use of the word 'intentionally.' The very same

effect also arises for people's use of 'intention,' 'deciding,' 'desire,' 'in favor of,' 'advocating,' and many other related expressions.

To get a grip on this phenomenon, it may be helpful to look in more detail at the actual procedure involved in conducting these studies. In one common experimental design, subjects are randomly assigned to receive either the story about harming the environment or the story about helping the environment and then, depending on the case, asked about the degree to which they agree or disagree with one of the following sentences:

(1) a. The chairman of the board harmed the environment intentionally.

○ ○ ○ ○ ○ ○ ○

definitely disagree — unsure — definitely agree

b. The chairman of the board helped the environment intentionally.

○ ○ ○ ○ ○ ○ ○

definitely disagree — unsure — definitely agree

When the study is conducted in this way, one finds that subjects show moderate agreement with the claim that the chairman harmed intentionally and moderate disagreement with the claim that he helped intentionally (Knobe 2004b). The difference between the ratings in these two conditions provides evidence that people's moral intuitions are affecting their intuitions about intentional action.

It appears, however, that this effect is not limited to the concept of intentional action specifically. For example, suppose we eliminate the word 'intentionally' and instead use the word 'decided.' The two sentences then become:

(2) a. The chairman decided to harm the environment.

b. The chairman decided to help the environment.

Faced with these revised sentences, subjects show more or less the same pattern of intuitions. They tend to agree with the claim that the agent decided to harm, while they tend to disagree with the claim that the agent decided to help (Pettit & Knobe forthcoming).

Now suppose we make the case a little bit more complex. Suppose we do not use the adverb 'intentionally' but instead use the verb 'intend.' So the sentences come out as:

(3) a. The chairman intended to harm the environment.

b. The chairman intended to help the environment.

One then finds a rather surprising result. People's responses in both conditions are shifted over quite far toward the 'disagree' side. In fact, people's intuitions end up being shifted over so far that they do not, on the whole, agree in either of the two conditions (Shepard 2009; cf. Cushman 2010; Knobe 2004a; McCann 2005). Nonetheless, the basic pattern of the responses remains the same. Even though people's responses don't go all the way over to the 'agree' side of the scale in either condition, they are still *more* inclined to agree in the harm case than they are in the help case.

Once one conceptualizes the issue in this way, it becomes possible to find an impact of moral considerations just about everywhere one looks. Take people's application of the concept *in favor*. Now consider a case in which an agent says:

> I know that this new procedure will [bring about some outcome]. But that is not what we should be concerned about. The new procedure will increase profits, and that should be our goal.

Will people say in such a case that the agent is 'in favor' of bringing about the outcome? Here again, it seems that moral judgments play a role. People disagree with the claim that the agent is 'in favor' when the outcome is morally good, whereas they stand at just about the midpoint between agreement and disagreement when the outcome is morally bad (Pettit & Knobe forthcoming). And similar effects have been observed for people's use of many other concepts: *desiring*, *intending*, *choosing*, and so forth (Pettit & Knobe forthcoming; Pettit & Knobe unpublished data; Tannenbaum et al. 2009).

Overall, these results suggest that the effect obtained for intuitions about intentional action is just one example of a far broader phenomenon. The effect does not appear to be limited to the concept *intentionally*, nor even to closely related concepts such as *intention* and *intending*. Rather, it seems that we are tapping into a much more general tendency whereby moral judgments impact the application of a whole range of different concepts used to pick out psychological states and processes.

2.3. Action trees

But the scope of the effect does not stop there. It seems also to apply to intuitions about the relations that obtain among the various actions an agent performs. Philosophers and cognitive scientists have often suggested that such relations could be represented in terms of an *action tree* (Goldman 1970; Mikhail 2007). Hence, the various actions performed by our chairman in the help case might be represented like this:

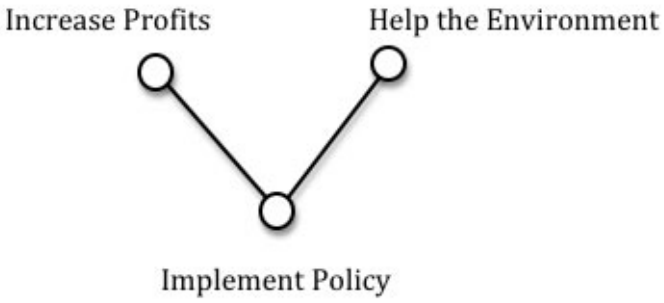

Figure 1: Action tree for the help case.

Needless to say, ordinary folks do not actually communicate with each other by writing out little diagrams like this one. Still, it seems that we can get a sense of how people are representing the action tree by looking at their use of various ordinary English expressions, e.g., by looking at the way they use the expressions 'in order to' and 'by.'

A number of complex issues arise here, but simplifying slightly, the key thing to keep in mind is that people only use 'in order to' for relations that go *upward* in the tree and they only use 'by' for relations that go *downward*. Thus, people are willing to say that the chairman 'implemented the program in order to increase profits' but not that he 'increased profits in order to implement the program.' And, conversely, they are willing to say that he 'increased profits by implementing the program' but not that he 'implemented the program by increasing profits.' Looking at people's intuitions about simple expressions like these, we can get a good sense of how they are representing the geometry of the action tree itself.

But now comes the tricky part. Experimental results indicate that people's intuitions about the proper use of these expressions can actually be influenced by their moral judgments (Knobe 2004; Knobe forthcoming). Hence, people are willing to say:

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

but not:

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

And, similarly, they are willing to say:

(3) The chairman increased profits by harming the environment.

but not:

(4) The chairman increased profits by helping the environment.

One natural way of explaining these asymmetries would be to suggest that people's moral judgments are having an effect on their representations of the action tree itself. For example, suppose that when people make a judgment that harming the environment is morally wrong, they thereby come to represent the corresponding node on the action tree as 'collapsing' into a lower node:

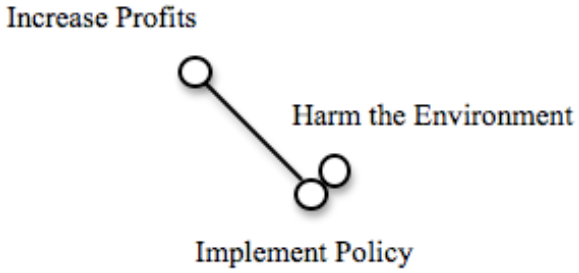

Figure 2: Action tree for the harm case.

The asymmetries we find for 'in order to' and 'by' would then follow immediately, without the need for any controversial assumptions about the semantics of these specific expressions. Although the issue here is a complex one, recent research does seem to be

supporting the claim that moral judgments are affecting action tree representations in this way (Knobe forthcoming; Ulatowski 2009).

2.4. Causation

All of the phenomena we have been discussing thus far may appear to be quite tightly related, and one might therefore suspect that the effect of morality would disappear as soon as one turns to other, rather different cases. That, however, seems not to be the case. Indeed, the very same effect arises in people's intuitions about *causation* (Alicke 2000; Cushman 2010; Hitchcock & Knobe forthcoming; Knobe forthcoming; Knobe & Fraser 2008; Solan & Darley 2001).

For a simple example here, consider the following vignette:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own.
>
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mailed them reminders that only administrators are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later, that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk.

Faced with this vignette, most subjects say that the professor did cause the problem but that the administrative assistant did not cause the problem (Knobe & Fraser 2008). Yet,

when we examine the case from a purely scientific standpoint, it seems that the professor's action and the administrative assistant's action bear precisely the same relation to the problem that eventually arose. The main difference between these two causal factors is just that the professor is doing something wrong (violating the departmental rule) while the administrative assistant is doing exactly what she is supposed to (acting in accordance with the rules of the department). So it appears that people's judgment that the professor is doing something wrong is somehow affecting their intuitions about whether or not the professor *caused* the events that followed.

Now, looking just at this one case, one might be tempted to suppose that the effect is not at all a matter of moral judgment but simply reflects people's intuitive sense that the professor's action is more 'unusual' or 'strange' than the administrative assistant's. But subsequent studies strongly suggest that there is something more afoot here. People continue to show the same basic effect even when they are informed that the administrative assistants *never* take pens whereas the professors always do (Roxborough & Cumby 2009), and there is a statistically significant effect whereby pro-life subjects are more inclined than pro-choice subjects to regard the act of seeking an abortion as a cause of subsequent outcomes (Cushman, Knobe, Sinnott-Armstrong 2008). All in all, the evidence seems strongly to suggest that people's moral judgments are actually impacting their causal intuitions.

2.5. Doing and allowing

People ordinarily distinguish between actually *breaking* something and merely *allowing* it to break, between actually *raising* something and merely *allowing* it to rise, between

actually *killing* someone and merely *allowing* a person to die. This distinction has come to be known as the distinction between *doing* and *allowing*.

To explore the relationship between people's intuitions about doing and allowing and their moral judgments, we used more or less the same methodology employed in these earlier studies (Cushman, et al. 2008). Subjects were randomly assigned to receive different vignettes. Subjects in one condition received a vignette in which the agent performs an action that appears to be morally permissible:

> Dr. Bennett is an emergency-room physician. An unconscious homeless man is brought in, and his identity is unknown. His organ systems have shut down and a nurse has hooked him up to a respirator. Without the respirator he would die. With the respirator and some attention from Dr. Bennett he would live for a week or two, but he would never regain consciousness and could not live longer than two weeks.
>
> Dr. Bennett thinks to himself, "This poor man deserves to die with dignity. He shouldn't spend his last days hooked up to such a horrible machine. The best thing to do would be to disconnect him from the machine."
>
> For just that reason, Dr. Bennett disconnects the homeless man from the respirator, and the man quickly dies.

These subjects were then asked whether it would be more appropriate to say that the doctor *ended* the homeless man's life or that he *allowed* the homeless man's life to end.

Meanwhile, subjects in the other condition were given a vignette that was almost exactly the same, except that the doctor's internal monologue takes a somewhat different turn:

> Dr. Bennett thinks to himself, "This bum deserves to die. He shouldn't sit here soaking up my valuable time and resources. The best thing to do would be to disconnect him from the machine."

These subjects were asked the same question: whether it would be more appropriate to say that the doctor ended the man's life or allowed it to end.

Notice that the doctor performs exactly the same behavior in these two vignettes, and in both vignettes, he performs this behavior in the hopes that it will bring about the man's death. The only difference between the cases lies in the moral character of the doctor's reasons for hoping that the man will die. Yet this moral difference led to a striking difference in people's intuitions about doing vs. allowing. Subjects who received the first vignette tended to say that the doctor 'allowed' the man's life to end, whereas subjects who received the second vignette tended to say that the doctor 'ended' the man's life. (Moreover, even within the first vignette, there was a correlation whereby subjects who thought that euthanasia was generally morally wrong were less inclined to classify the act as an 'allowing.') Overall, then, the results of the study suggest that people's moral judgments are influencing their intuitions here as well.

It would, of course, be foolhardy to draw any very general conclusions from this one study, but the very same effect has also been observed in other studies using quite different methodologies (Cushman et al. 2008), and there is now at least some good provisional evidence in support of the view that people's intuitions about doing and allowing can actually be influenced by their moral judgments.

2.6. Additional effects

Here we have discussed just a smattering of different ways in which people's moral judgments can impact their intuitions about apparently non-moral questions. But our review has been far from exhaustive: there are also studies showing that moral judgments can affect intuitions about *knowledge* (Beebe & Buckwalter forthcoming), *happiness* (Nyholm 2009), *valuing* (Knobe & Roedder forthcoming), *act individuation* (Ulatowski 2009), *freedom* (Phillips & Knobe 2009), and *naturalness* (Martin 2009). Given that all of these studies were conducted just in the past few years, it seems highly probable that a number of additional effects along the same basic lines will emerge in the years to come.

3. Alternative explanations

Thus far, we have seen that people's ordinary application of a variety of different concepts can be influenced by moral considerations. The key question now is how to explain this effect. Here we face a choice between two basic approaches. One approach would be to suggest that moral considerations actually figure in the fundamental competencies people use to understand the world. The other would be to adopt what I will call an *alternative explanation*. That is, one could suggest that moral considerations play no role at all in the underlying competencies but that certain additional factors are somehow 'biasing' or 'distorting' people's cognitive processes and thereby allowing their intuitions to be affected by moral judgments.

The first thing to notice about the debate between these two approaches is that we are unlikely to make much progress on it as long as the two positions are described only in these abstract, programmatic terms. Thus suppose that we are discussing a new experimental result, and someone says: 'Well, it could always turn out that this effect is

due to some kind of interfering factor.' How would we even begin to test such a conjecture? As long as the claim is just about the possibility of 'some kind of interfering factor,' it is hard to know where one could go to look for confirming or disconfirming evidence.

Fortunately, however, the defenders of alternative hypotheses have not simply put forward these sorts of abstract, programmatic conjectures. Instead, they have developed sophisticated models that make it possible to offer detailed explanations of the available experimental data. Such models start out with the idea that people's actual competence includes no role for moral considerations, but they then posit various additional psychological factors that explain how people's moral judgments might nonetheless influence their intuitions in specific cases. Each such alternative explanation then generates further predictions, which can in turn be subjected to experimental test. There has been a great deal of research in recent years devoted to testing these models, including some ingenious new experiments that enable one to get a better handle on the complex cognitive processes underlying people's intuitions. At this point, then, the best approach is probably just to look in detail at some of the most prominent explanations that have actually been proposed and the various experiments that have been devised to test them.

3.1. The motivational bias hypothesis

Think of the way a District Attorney's office might conduct its business. The DA decides to prosecute a suspect and hands the task over to a team of lawyers. These lawyers then begin looking at the case. Presumably, though, they do not examine the evidence with

perfectly unbiased eyes. They have been hired to secure a conviction, and they are looking at the evidence with a view to achieving this goal (cf. Tetlock 2002). One might say that they are under the influence of a *motivational bias*.

A number of researchers have suggested that a similar mechanism might be at the root of the effects we have been discussing here (Alicke 2008; Nadelhoffer 2006a). Perhaps people just read through the story and rapidly and automatically conclude that the agent is to blame. Then, after they have already reached this conclusion, they begin casting about for ways to justify it. They try to attribute anything they can – intention, causation, etc. – that will help to justify the blame they have already assigned. In essence, the suggestion is that the phenomena under discussion here can be understood as the results of a motivational bias.

This suggestion would involve a reversal of the usual view about the relationship between people's blame judgments and their intuitions about intention, causation, and so forth. The usual view is that this relationship looks something like this:

Figure 3: Traditional account of the process underlying blame ascription.

Here, the idea is that people first determine that the agent fulfilled the usual criteria for moral responsibility (intention, cause, etc.) and then, on the basis of this initial judgment, go on to determine that the agent deserves blame. This sort of model has a strong intuitive appeal, but it does not seem capable of explaining the experimental data

reviewed above. After all, if people determine whether or not the agent caused the outcome before they make any sort of moral judgment, how could it be that their moral judgments affect their intuitions about causation?

To resolve this question, one might develop a model that goes more like this:

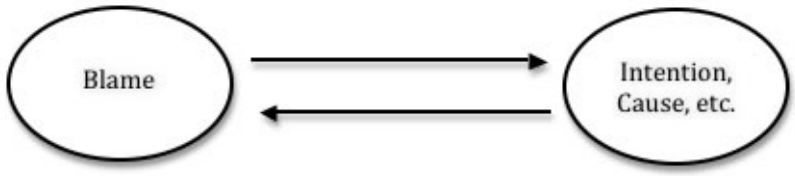

Figure 4: Motivational bias account of blame ascription.

In this revised model, there is a reciprocal relationship between people's blame judgments and their intuitions about intention, causation, etc. As soon as people observe behavior of a certain type, they become motivated to find some way of blaming the agent. They then look to the evidence and try to find a plausible argument in favor of the view that the agent fulfills all of the usual criteria for responsibility. If they can construct a plausible argument there, they immediately blame the agent. Otherwise, they reluctantly determine that the agent was not actually blameworthy after all. In short, the hypothesis says that people's intuitions about intention and causation affect their blame judgments but that the causal arrow can also go in the other direction, with people's drive to blame the agent distorting their intuitions about intention and causation.

One of the main sources of support for such a hypothesis is the well-established body of theoretical and experimental work within social psychology exploring similar effects in other domains. There is now overwhelming evidence that motivational biases can indeed lead people to interpret evidence in a biased manner (for a review, see Kunda

1990), and within moral psychology specifically, there is a growing body of evidence suggesting that people often adopt certain views as part of a post-hoc attempt to justify prior moral intuitions (Haidt 2001; Ditto et al. 2009). So the motivational bias hypothesis is perhaps best understood as the application to a new domain of a theoretical perspective that is already quite well supported elsewhere.

More importantly, the hypothesis makes it possible to explain all of the existing results without supposing that moral considerations actually play any role at all in any of the relevant competencies. The thought is that people's competencies are entirely non-moral but that a motivational bias then interferes with our ability to apply these concepts correctly. (An analogous case: If John sleeps with Bill's girlfriend, Bill may end up concluding that John's poetry was never really any good – but that does not mean that Bill's underlying criteria for poetry actually involve any reference to sexual behavior.)

All in all, then, what we have here is an excellent hypothesis. It draws on well-established psychological theory, provides a clear explanation of existing results, and offers a wealth of new empirically testable predictions. The one problem is that when researchers actually went out and tested those new predictions, none of them were empirically confirmed. Instead, the experimental results again and again seemed to go against what would have been predicted on the motivational bias view. At this point, the vast majority of researchers working on these questions have therefore concluded that the motivational bias hypothesis cannot explain the full range of experimental findings and that some other sort of psychological process must be at work here (Hindriks forthcoming; Machery forthcoming; McCann 2005; Nichols & Ulatowski 2007; Turner 2004; Wright & Bengson 2009; Young et al. 2006).

3.1.1. The usual way of understanding the motivational bias hypothesis is that reading through certain kinds of vignettes triggers an immediate affective reaction, which then distorts people's subsequent reasoning (Nadelhoffer 2006a). An obvious methodology for testing the hypothesis is therefore to find people who *don't* have these immediate affective reactions and then check to see whether these people still show the usual effect.

Young, Cushman, Adolphs, Tranel and Hauser (2006) did just that. They took the very same cases we discussed above and gave these cases to subjects who had lesions in the ventromedial prefrontal cortex (VMPFC). Previous experiments had shown that such subjects have massive deficits in the ordinary capacity for affective response. They show little or no affective response in cases where normal subjects would respond strongly (Damasio et al. 1990), and when they are presented with moral dilemmas in which most people's answers seem to be shaped by affective responses, they end up giving answers that are radically different from those given by normal subjects (e.g., Koenigs, Young et al. 2007). The big question was whether they would also give unusual answers on the types of questions we have been examining here.

The results showed that they did not (Young et al. 2006). Just like normal subjects, the VMPFC patients said that the chairman harmed the environment intentionally but helped the environment unintentionally. In fact, *one hundred percent* of patients in this study said that the environmental harm was intentional. On the basis of this experimental result, Young and colleagues concluded that the asymmetry observed in normal subjects was not, in fact, due to an affective reaction.

But, of course, even if it turns out that affective reactions play no role in these effects, the motivational bias hypothesis would not necessarily be refuted (Alicke 2008). After all, it is important to distinguish carefully between affect and motivation, and we need to acknowledge the possibility that people are experiencing a motivational bias that does not involve any kind of affect at all. Perhaps people just calmly observe certain behaviors, rapidly arrive at certain moral appraisals, and then find themselves trying to justify a judgment of blame.

This proposal is, I believe, an interesting and suggestive one. To address it properly, we will need to develop a more complex theoretical framework.

3.1.2. To begin with, we need to distinguish between a variety of different types of moral judgment. One type of moral judgment is a judgment of *blame*. This is the type of judgment we have been discussing thus far, and it certainly does play an important role in people's psychology. But it is not the only type of moral judgment people make. They also make judgments about whether an agent did something morally *wrong*, about whether a behavior violated people's moral *rights*, about whether its consequences were *bad*. A complete theory of moral cognition would have to distinguish carefully between these various types of moral judgments and explain how each relates to people's intuitions about intention, causation, etc.

In any case, as soon as we distinguish these various types of moral judgment, we see that it would be possible for people's intuitions to be influenced by their moral judgments even if these intuitions are not influenced by *blame* in particular. In fact, a

growing body of experimental evidence suggests that the process actually works something like this:

Figure 5: Distinct processes of moral judgment.

This model involves a quite radical rejection of the view that people's intuitions about intention, causation, etc. are distorted by judgments of blame. Not only are these intuitions not *distorted* by blame, they are not even influenced by blame at all. Rather, people start out by making some other type of moral judgment, which then influences their intuitions about intention and causation, which in turn serves as input to the process of assessing blame.

Though this model may at first seem counterintuitive, it has received support from experimental studies using a wide variety of methodologies. To take one example, Guglielmo and Malle (2009a) gave subjects the vignette about the chairman and the environment and then used structural equation modeling to test various hypotheses about the relations among the observed variables. The results did not support a model in which blame judgments affected intuitions about intentional action. In fact, the analysis supported a causal model that went in precisely the opposite direction: it seems that people are first arriving at an intuition about intentional action and that this intuition is then impacting their blame judgments. In short, whatever judgment it is that affects

people's intentional action intuitions, the statistical results suggest that it is not a judgment of blame per se.

In a separate experiment, Guglielmo and Malle (2009b) used reaction time measures to determine how long it took subjects to make a variety of different types of judgments. The results showed that people generally made judgments of intentional action *before* they made judgments of blame. (There was even a significant effect in this direction for some, though not all, of the specific cases we have been considering here.) But if the blame judgment does not even take place until after the intentional action judgment has been completed, it seems that people's intentional action judgments cannot be distorted by feedback from blame.

Finally, Keys and Pizarro (unpublished data) developed a method that allowed them to manipulate blame and then look for an effect on intuitions about intentional action. Subjects were given the vignettes about the agent who either helps or harms the environment, but they were also randomly assigned to receive different kinds of information about the character of this agent. Some were given information that made agent look like a generally nice person; others were given information that made the agent look like a generally nasty person. The researchers could then examine the impact of this manipulation on intuitions about blame and about intentional action. Unsurprisingly, people's intuitions about blame were affected by the information they received about the agent's character, but – and this is the key result of the experiment – this information had no significant impact on people's intuitions about intentional action. Instead, intuitions about intentional action were affected only by information about the actual behavior (helping vs. harming) the agent was said to have performed.[3]

In the face of these new results, friends of the motivational bias view might simply to retreat to a weaker position. They might say: 'Okay, so we initially suggested that people's intuitions were distorted by an affective reaction associated with an impulse to blame, but we now see that the effect is not driven by affect and is not caused specifically by blame. Still, the basic idea behind the theory could nonetheless be on track. That is to say, it could still be that people's intuitions are being distorted by an effort to justify some kind of moral judgment…'

3.1.3. This approach certainly sounds good in the abstract, but as one proceeds to look carefully at the patterns of intuition observed in specific cases, it starts to seem less and less plausible. The difficulty is that the actual patterns observed in these cases just don't make any sense as an attempt to justify prior moral judgments.

For a simple example, consider the case in which the receptionist runs out of pens and people conclude that the professor is the sole cause of the problem that results. In this case, it seems that some kind of moral judgment is influencing people's intuitions about causation, but which moral judgment is doing the work here? One obvious hypothesis would be that people's intuitions about causations are being influenced by a judgment that *the agent deserves blame for the outcome*. If this hypothesis were correct, it would make a lot of sense to suggest that people's intuitions were being distorted by a motivational bias. The idea would be that people want to conclude that the professor is to blame for a particular outcome and, to justify this conclusion, they say that he is the sole cause of this outcome.

The one problem is that this hypothesis does not actually appear to be correct. It does not seem that people's causal intuitions truly are being influenced by a judgment that the agent is to blame for the outcome. Instead, the data suggest that people's intuitions are simply being influenced by a judgment that *the agent's action itself is bad*. So, for example, in the case at hand, we can distinguish two different moral judgments that people might make:

1. The professor is to blame for the outcome (the receptionist's lack of pens).
2. There is something bad about the professor's action (taking a pen from the desk).

The key claim now is that it is the second of these judgments, rather than the first, that is influencing people's intuition that the professor caused outcome.

To test this claim empirically, we need to come up with a case in which the agent is judged to have performed a bad action but in which the agent is nonetheless not judged to be blameworthy for the outcome that results. One way to construct such a case would be to modify our original story by switching the outcome over to something *good*. (For example: the receptionist was planning to stab the department chair's eye out with a pen, but now that all of the pens have been taken, her plan is thwarted, and the department chair's eyes are saved.) In such a case, the professor would still be performing a bad action, but there would not even be a question as to whether he was 'to blame' for the outcome that resulted, since there would be no bad outcome for which anyone could deserve blame.

Experiments using this basic structure have arrived at a surprising pattern of results (Hitchcock & Knobe forthcoming). Even when the outcome has been switched to something good, people continue to have the same causal intuitions. They still conclude

that the agent who performed the bad action is more of a cause than the agent who performed the good action. Yet when the outcome is something good, it seems impossible to explain this pattern in terms of a motivational bias. After all, friends of the motivational bias hypothesis would then have to say that people are displeased with the agent who performs the bad action, that their intuitions thereby become distorted by moral judgment, and that they end up being motivated to conclude: 'This bad guy must have been the sole cause of the wonderful outcome that resulted.' It seems quite difficult, however, to see how such a conclusion could possibly serve as a post-hoc justification for some kind of negative moral judgment.

3.1.4. Of course, it might ultimately prove possible to wriggle out of all of these difficulties and show that the data we have amassed here does not refute the motivational bias hypothesis. But even then, a larger problem would still remain. This problem is that no one ever seems to be able to produce any positive evidence in favor of the hypothesis. That is, no one seems to be able to provide evidence that motivational biases are at the root of the particular effects under discussion here. There is, of course, plenty of evidence that motivational biases do in general exist (e.g., Kunda 1990), and there are beautiful experimental results showing the influence of motivational biases in other aspects of moral cognition (Alicke 2000; Haidt 2001; Ditto, Pizarro & Tannenbaum forthcoming), but when it comes to the specific effects under discussion here, there are no such experiments. Instead, the argument always comes down to something like: 'This explanation turned out to be true for so many other effects, so it is probably true for these as well.'

It now appears that this strategy may have been leading us astray. The basic concepts at work in the motivational bias explanation – affective reactions, post-hoc rationalization, motivated reasoning – have proved extraordinarily helpful in understanding other aspects of moral cognition. But moral cognition is a heterogeneous phenomenon. What proves helpful in thinking about certain aspects of it may prove utterly irrelevant in thinking about others.

3.2. The conversational pragmatics hypothesis

Let us turn, then, to a second possible alternative hypothesis. When people are engaged in ordinary discussions, their use of words does not simply serve as a straightforward reflection of the way they apply their underlying concepts. Instead, people strive to act as helpful conversation partners, following certain complex principles that enable them to provide useful information to their audience. The study of these principles falls under the heading of 'conversational pragmatics,' and researchers engaged in this study have illuminated many puzzling aspects of the way people ordinarily use language in communication. A number of researchers have suggested that this approach might also serve to explain the phenomena we are trying to understand here (Adams & Steadman 2004a, 2004b; Driver 2008a, 2008b).

To get a sense for this hypothesis, it might be helpful to start out by looking at a potentially analogous case in another domain. Imagine that you have a bathroom in your building but that this bathroom is completely non-functional and has been boarded up for the past three years. And now imagine that someone hands you a questionnaire that asks:

> Do you have a bathroom in your building?

__ Yes __ No

It does seem that your underlying concept *bathroom* might correctly apply to the room in your building, but when you receive this question, you immediately have an understanding of what the questioner really wants to know – namely, whether or not you have a bathroom that actually works — and you might therefore choose to check the box marked 'No.'

With these thoughts in mind, consider what might happen when subjects receive a questionnaire that asks whether they agree or disagree with the sentence:

The chairman of the board harmed the environment intentionally.

○ ○ ○ ○ ○ ○ ○

definitely disagree — unsure — definitely agree

It might be thought that people's underlying concept *intentional* does not, in fact, apply to cases like this one but that, as soon as they receive the questionnaire, they form an understanding of what the questioner really wants to know. The real question here, they might think, is whether the chairman deserves to be blamed for his behavior, and they might therefore check the circle marked 'definitely agree.'

Similar remarks might be applied to many of the other effects described above. Thus, suppose that subjects are asked whether they agree or disagree with the sentence:

The administrative assistant caused the problem.

○ ○ ○ ○ ○ ○ ○

definitely disagree — unsure — definitely agree

It might be thought that people's concept *cause* does apply in cases like this one, but it also seems that subjects might quite reasonably infer that the real point of the question is to figure out whether the administrative assistant deserves blame for this outcome and that they might therefore check the circle marked 'definitely disagree.'

Before going on any farther, it might be helpful to take a moment to emphasize just how different this pragmatic hypothesis is from the motivational bias hypothesis we discussed above. The motivational bias hypothesis posits an error that affects people's understanding of certain morally relevant events. By contrast, the pragmatic hypothesis does not involve any error or even any effect on people's understanding of events. It simply suggests that people are applying certain kinds of conversational rules. The basic idea is that moral considerations aren't actually affecting people's fundamental understanding of the situation; it's just that moral considerations do sometimes affect people's view about which particular words would be best used to describe it.

In any case, although the two hypotheses are very different in their theoretical approaches, they have proved remarkably similar in their ultimate fate. Like the motivational bias hypothesis, the pragmatic hypothesis initially looked very promising – a clear and plausible explanation, backed by a well-supported theoretical framework – but, as it happened, the actual empirical data just never came out the way the pragmatic hypothesis would predict. Indeed, the pragmatic hypothesis suffers from many of the same problems that plagued the motivational bias hypothesis, along with a few additional ones that are all its own.

3.2.1. One way to test the hypothesis would be to identify subjects who show an inability to use conversational pragmatics in the normal way and then to check to see whether these subjects still show the usual effect. Zalla, Machery and Leboyer (2010) did exactly that. They took the story about the chairman who harms or helps the environment and presented it to subjects with Asperger's syndrome, a developmental disorder characterized by difficulties in certain forms of communication and a striking inability to interact normally with others. Previous studies had shown that subjects with Asperger's display remarkable deficits in the capacity to understand conversational pragmatics, tending instead to answer questions in the most literal possible way (e.g., De Villiers, Stainton & Szatmari 2006; Surian et al. 1996). If the original effect had been due entirely to pragmatic processes, one might therefore have expected subjects with Asperger's to respond quite differently from typically developing subjects.

But that is not what Zalla and colleagues found. Instead, they found that subjects with Asperger's showed exactly the same pattern of responses that typically developing subjects did. Just like typically developing subjects, they tended to say that the chairman harmed the environment intentionally but helped it unintentionally. This result suggests that the pattern displayed by typically developing subjects is not, in fact, a product of their mastery of complex pragmatic principles.

3.2.2. Of course, the study of linguistic deficits in people with Asperger's brings up a host of complex issues, and this one experiment certainly should not be regarded as decisive. The thing to notice, though, is that results from a variety of other tests point toward the same basic conclusion, offering converging evidence for the claim that the

effect here is not a purely pragmatic one (Adams & Steadman 2007; Knobe 2004; Nichols & Ulatowski 2007; for a review, see Nadelhoffer 2006b).

Indeed, one can obtain evidence for this claim using one of the oldest and most widely known tests in the pragmatics literature. Recall that we began our discussion of conversational pragmatics with a simple example. If a person says 'There is a bathroom in the building,' it would be natural to infer that this bathroom is actually in working order. But now suppose that we make our example just a little bit more complex. Suppose that the person utters two sentences: 'There is a bathroom in the building. However it is not in working order.' Here it seems that the first sentence carries with it a certain sort of pragmatic significance but that the second sentence then eliminates the significance that this first sentence might otherwise have had. The usual way of describing this phenomenon is to say that the pragmatic implicatures of the first sentence have been *cancelled* by the second (Grice 1989).

Using this device of cancellation, we could then construct a questionnaire that truly would accurately get at people's actual concept of bathrooms. For example, subjects could be asked to select from among the options:

__ There is no bathroom in the building.

__ There is a bathroom in the building, and it is in working order.

__ There is a bathroom in the building, but it is not in working order.

Subjects could then feel free to signify the presence of the bathroom by selecting the third option, secure in the knowledge that they would not thereby be misleadingly conveying an impression that the bathroom actually did work.

In a recent experimental study, Nichols and Ulatowski (2007) used this same approach to get at the impact of pragmatic factors in intuitions about intentional action. Subjects were asked to select from among the options:

__ The chairman *intentionally* harmed the environment, and he is responsible for it

__ The chairman didn't *intentionally* harm the environment, but he is responsible for it.

As it happened, Nichols and Ulatowski themselves believed that the original effect was entirely pragmatic, and they therefore predicted that subjects would indicate that the behavior was unintentional when they had the opportunity to do so without conveying the impression that the chairman was not to blame. But that is not at all how the data actually came out. Instead, subjects were just as inclined to say that the chairman acted intentionally in this new experiment as they were in the original version. In light of these results, Nichols and Ulatowski concluded that the effect was not due to pragmatics after all.

3.2.3. Finally, there is the worry that, even if conversational pragmatics might provide a somewhat plausible explanation of some of the effects described above, there are other effects that it cannot explain at all. Hence, the theory of conversational pragmatics would fail to explain the fact that moral considerations exert such a pervasive effect on a wide range of different kinds of judgments.

The pragmatic hypothesis was originally proposed as an explanation for people's tendency to agree with sentences like:

The chairman of the board harmed the environment intentionally.

And when the hypothesis is applied to cases like this one, it does look at least initially plausible. After all, it certainly does seem that a sentence like 'He did not harm the environment intentionally' could be used to indicate that the agent was not, in fact, to blame for his behavior.

But now suppose we take that very same hypothesis and apply it to sentences like:

The chairman harmed the environment in order to increase profits.

Here the hypothesis does not even begin to get a grip. There simply isn't any conversational rule according to which one can indicate that the chairman is not to blame by saying something like: 'He didn't do that in order to increase profits.' No one who heard a subject uttering such a sentence would ever leave with the impression that it was intended as a way of exculpating or excusing the chairman.

Of course, one could simply say that the pragmatics hypothesis does explain the effect on 'intentionally' but does not explain the corresponding effect on 'in order to.' But such a response would take away much of the motivation for adopting the pragmatics hypothesis in the first place. The hypothesis was supposed to give us a way of explaining how moral considerations could impact people's use of certain words without giving up on the idea that people's underlying concepts were entirely morally neutral. If we now accept a non-pragmatic explanation of the effect for 'in order to,' there is little reason not to accept a similar account for 'intentionally' as well.

3.3. Summary

Looking through these various experiments, one gradually gets a general sense of what has been going wrong with the alternative explanations. At the core of these

explanations is the idea that people start out with an entirely non-moral competence but that some additional factor then interferes and allows people's actual intuitions to be influenced by moral considerations. Each alternative explanation posits a different interfering factor, and each explanation thereby predicts that the whole effect will go away if this factor is eliminated. So one alternative explanation might predict that the effect will go away when we eliminate a certain emotional response, another that it will go away when we eliminate certain pragmatic pressures, and so forth.

The big problem is that these predictions never actually seem to be borne out. No one has yet found a way of eliminating the purported interfering factors and thereby making the effect go away. Instead, the effect seems always to stubbornly reemerge, coming back again and again despite all our best efforts to eliminate it.

Now, one possible response to these difficulties would be to suggest that we just need to try harder. Perhaps the relevant interfering factor is an especially tricky or well-hidden one, or maybe there are a whole constellation of different factors in place here, all working together to generate the effects observed in the experiments. When we finally succeed in identifying all of the relevant factors, we might be able to find a way of eliminating them all and thereby allowing people's purely non-moral competence to shine through unhindered.

Of course, it is at least possible that such a research program would eventually succeed, but I think the most promising approach at this point would be to try looking elsewhere. In my view, the best guess about why no one has been able to eliminate the interfering factors is that there just *aren't* any such factors. It is simply a mistake to try to understand these experimental results in terms of a purely non-moral competence which

then gets somehow derailed by various additional factors. Rather, the influence of moral considerations that comes out in the experimental results truly is showing us something fundamental about the nature of the basic competencies people use to understand their world.

4. Competence theories

Let us now try to approach the problem from a different angle. Instead of focusing on the interfering factors, we will try looking at the competence itself. The aim will be to show that something about the very nature of this competence is allowing people's moral judgments to influence their intuitions.

4.1. General approach

At the core of the approach is a simple and straightforward assumption that has already played an enormously important role in numerous fields of cognitive science. Specifically, I will be relying heavily on the claim that we make sense of the things that actually happen by considering *other ways things might have been* (Byrne 2005; Kahneman & Miller 1986; Roese 1997).

A quick example will help to bring out the basic idea here. Suppose that we come upon a car that has a dent in it. We might immediately think about how the car would have looked if it did not have this dent. Thus, we come to understand the way the car actually is by considering another way that it could have been and comparing its actual status to this imagined alternative.

An essential aspect of this process, of course, lies in our ability to select among all the possible alternatives just the few that prove especially relevant. Hence, in the case at hand, we would immediately consider the possibility that the car could have been undented and think: 'Notice that this car is dented rather than undented.' But then there are all sorts of other alternatives that we would immediately reject as irrelevant or not worth thinking about. We would not take the time, e.g., to consider the possibility that the car could have been levitating in the air and then think: 'Notice that the car is standing on the ground rather than levitating in the air.'

Our ability to pick out just certain specific alternatives and ignore others is widely regarded as a deeply important aspect of human cognition, which shapes our whole way of understanding the objects we observe. It is, for example, a deeply important fact about our way of understanding the dented car that we compare it to an undented car. If we had instead compared it to a levitating car, we would end up thinking about it in a radically different way.

A question now arises as to why people focus on certain particular alternative possibilities and ignore others. The answer, of course, is that all sorts of different factors can play a role here. People's selection of specific alternative possibilities can be influenced by their judgments about controllability, about recency, about statistical frequency, about non-moral forms of goodness and badness (for reviews, see Byrne 2005; Kahneman & Miller 1986; Roese 1997). But there is also another factor at work here that has not received quite as much discussion in the existing literature. A number of studies have shown that people's selection of alternative possibilities can be influenced by their *moral judgments* (McCloy & Byrne 2000; N'gbala & Branscombe 1995). In other words,

people's intuition about which possibilities are relevant can be influenced by their judgments about which actions are morally right.

For a simple illustration, take the case of the chairman who hears that he will be helping the environment but reacts with complete indifference. As soon as one hears this case, one's attention is drawn to a particular alternative possibility:

(1) Notice that the chairman reacted in this way, rather than specifically preferring that the environment be helped.

This alternative possibility seems somehow to be especially relevant, more relevant at least than many other possibilities we could easily imagine. In particular, one would not think:

(2) Notice that the chairman reacted in this way rather than specifically trying to avoid anything that would help the environment.

Of course, one could imagine the chairman having this latter sort of attitude. One could imagine him saying: 'I don't care at all whether we make profits. What I really want is just to make sure that the environment is harmed, and since this program will help the environment, I'm going to do everything I can to avoid implementing it.' Yet this possibility has a kind of peculiar status. It seems somehow preposterous, not even worth considering. But why? The suggestion now is that moral considerations are playing a role in people's way of thinking about alternative possibilities. Very roughly, people regard certain possibilities as relevant because they take those possibilities to be especially good or right.

With these thoughts in mind, we can now offer a new explanation for the impact of moral judgments on people's intuitions. The basic idea is just that people's intuitions

in all of the domains we have been discussing – causation, doing/allowing, intentional action, and so on – rely on a comparison between the actual world and certain alternative possibilities. Since people's moral judgments influence the selection of alternative possibilities, these moral judgments end up having a pervasive impact on the way people make sense of human beings and their actions.[4]

4.2. A case study

To truly spell out this explanation in detail, one would have to go through each of the different effects described above and show how each of these effects can be explained on a model in which moral considerations are impacting people's way of thinking about alternative possibilities. This would be a very complex task, and we will not attempt it here. Let us proceed instead by picking just one concept whose use appears to be affected by moral considerations. We can then offer a model of the competence underlying that one concept and thereby illustrate the basic approach. For these illustrative purposes, let us focus on the concept *in favor*.

We begin by introducing a fundamental assumption that will guide the discussion that follows. The assumption is that people's representation of the agent's attitude is best understood, not in terms of a simple dichotomy between 'in favor' and 'not in favor,' but rather in terms of a whole *continuum* of different attitudes an agent might hold.[5] So we will be assuming that people can represent the agent as strongly in favor, as strongly opposed, or as occupying any of the various positions in between. For simplicity, we can depict this continuum in terms of a scale running from *con* to *pro*.

Figure 6: Continuum of attitude ascription.

Looking at this scale, it seems that an agent whose attitude falls way over on the *con* side will immediately be classified as 'not in favor' and that an agent whose attitude falls way over on the *pro* side will immediately be classified as 'in favor.' But now, of course, we face a further question. How do people determine the threshold at which an agent's attitude passes over from the category 'not in favor' to the category 'in favor'?

To address this question, we will need to add an additional element to our conceptual framework. Let us say that people assess the various positions along the continuum by comparing each of these positions to a particular sort of alternative possibility. We can refer to this alternative possibility as the *default*. Then we can suggest that an agent will be counted as 'in favor' when his or her attitude falls sufficiently far beyond this default point.

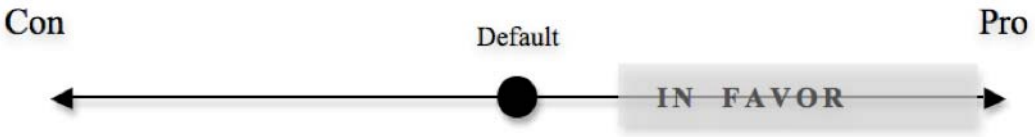

Figure 7: Criteria for ascription of 'in favor.'

The key thing to notice about this picture is that there needn't be any single absolute position on the continuum that always serves as the threshold for counting an agent as 'in favor.' Instead, the threshold might vary freely depending on which point gets picked out as the default.

To get a sense for the idea at work here, it may be helpful to consider a closely analogous problem. Think of the process a teacher might use in assigning grades to students. She starts out with a whole continuum of different percentage scores on a test, and now she needs to find a way to pick out a threshold beyond which a given score will count as an A. One way to do this would be to introduce a general rule, such as 'a score always counts as an A when it is at least 20 points above the default.' Then she can pick out different scores as the default on different tests – treating 75% as default on easy tests, 65% as default on more difficult ones – and the threshold for counting as an A will vary accordingly.

The suggestion now is that people's way of thinking about attitudes uses this same sort of process. People always count an agent as 'in favor' when his or her attitude falls sufficiently far beyond the default, but there is no single point along the continuum that is treated as default in all cases. Different attitudes can be treated as default in different cases, and the threshold for counting as 'in favor' then shifts around from one case to the next.

Now we arrive at the crux of the explanation. The central claim will be that people's moral judgments affect their intuitions *by shifting the position of the default*. For morally good actions, the default is to have some sort of pro-attitude, whereas for morally bad actions, the default is to have some sort of con-attitude. The criteria for 'in favor' then vary accordingly.

Suppose we now apply this general framework to the specific vignettes used in the experimental studies. When it comes to helping the environment, it seems that the default attitude is a little bit toward the *pro* side. That is to say, the default in this case is

to have at least a slightly positive attitude – not necessarily a deep or passionate attachment, but at least some minimal sense that helping the environment would be a nice thing to do. An attitude will then count as 'in favor' to the extent that it goes sufficiently far beyond this default point.

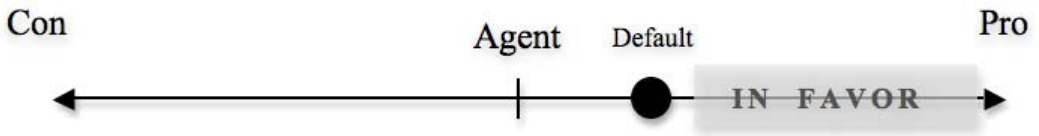

Figure 8: Representation of the continuum for the help case.

But look at the position of the agent's actual attitude along this continuum. The agent is not even close to reaching up to the critical threshold here – he is only interested in helping the environment as a side-effect of some other policy – and people should therefore conclude that he does not count as 'in favor' of helping.

Now suppose we switch over to the harm case. There, we find that the agent's actual attitude has remained constant, but the default has changed radically. When it comes to harming the environment, the default is to be at least slightly toward the *con* side – not necessarily showing any kind of vehement opposition, but at least having some recognition that harming the environment is a bad thing to do. An agent will then count as 'in favor' to the extent that his attitude goes sufficiently far beyond this default.

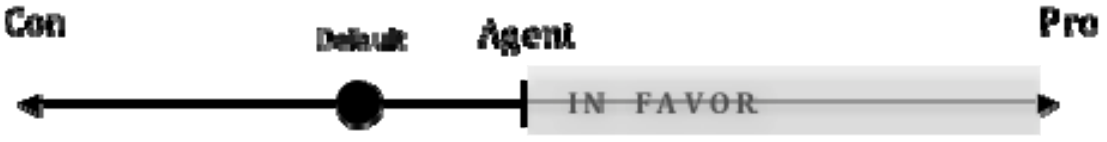

Figure 9: Representation of the continuum for the harm case.

In this new representation, the agent's actual attitude remains at exactly the same point it was above, but its position relative to the default is now quite different. This time, the agent falls just about at the critical threshold for counting as 'in favor,' and people should therefore be just about at the midpoint in their intuitions as to whether he was in favor of harming – which, in fact, is exactly what the experimental results show.

Notice how sharply this account differs from the alternative hypotheses discussed above. On those alternative hypotheses, people see that the agent harmed the environment, want to blame him for his behavior, and this interest in blame then shapes the way they conceptualize or describe various aspects of the case. The present account says nothing of the kind. Indeed, the account makes no mention at all of blame. Instead, it posits a role for an entirely different kind of moral judgment – a judgment that could be made even in the absence of any information about this specific agent or his behaviors. The claim is that before people even begin considering what actually happened in the case at hand, they can look at the act of harming the environment and make a judgment about what sort of attitude an agent could be expected to hold toward it. This judgment then serves as a standard which they can use to make sense of the behavior they actually observe.

4.3. Extending the model

What we have here is a model of the competence underlying people's use of one particular concept. The key question now is whether this same basic approach can be applied to the various other concepts discussed above. In a series of recent papers, I have argued that it can be used to explain the impact of moral judgment on people's intuitions about freedom, knowledge and causation[6] (Hitchcock & Knobe forthcoming; Pettit & Knobe forthcoming; Phillips & Knobe 2009), but new studies are coming out all the time,

and we may soon be faced with experimental results that the model cannot explain. At any rate, one certainly should not expect that this model will turn out to be correct in every detail. Presumably, further work will show that it needs to be revised or expanded in various ways, and perhaps it will even have to be scrapped altogether.

In the present context, however, our concern is not so much to explore the details of this one model as to use it as a way of illustrating a more general approach and the contrast between this approach and the one we saw in the alternative explanations described above. The alternative explanations start out with the idea that the relevant competencies are entirely non-moral but that some additional factor then interferes and allows people's intuitions to be influenced by moral considerations. These explanations therefore predict that it should be possible, at least in principle, to eliminate the interfering factors and examine the judgments people make in the absence of this influence. By contrast, in the approach under discussion here, moral considerations are not understood as some kind of extra factor that gets added in on top of everything else. Instead, the whole process is suffused with moral considerations from the very beginning. Hence, in this approach, no real sense can be attached to the idea of eliminating the role of morality and just watching the basic process unfold in its pure, non-moral form.

5. Conclusion

This paper began with a metaphor. The suggestion was that people's ordinary way of making sense of the world might be similar, at least in certain respects, to the way research is conducted in a typical modern university. Just as a university would have specific departments devoted especially to the sciences, our minds might include certain

specific psychological processes devoted especially to constructing a roughly 'scientific' kind of understanding.

If one thinks of the matter in this way, one immediately arrives at a certain picture of the role of moral judgments in people's understanding as a whole. In a university, there might be faculty members in the philosophy department who were hired specifically to work on moral questions, but researchers in the sciences typically leave such questions to one side. So maybe the mind works in much the same way. We might have certain psychological processes devoted to making moral judgments, but there would be other processes that focus on developing a purely 'scientific' understanding of what is going on in a situation and remain neutral on all questions of morality.

I have argued that this picture is deeply mistaken. The evidence simply does not suggest that there is a clear division whereby certain psychological processes are devoted to moral questions and others are devoted to purely scientific questions. Instead, it appears that everything is jumbled together. Even the processes that look most 'scientific' actually take moral considerations into account. It seems that we are moralizing creatures through and through.

Notes:

[1] For comments on earlier drafts, I am deeply grateful to John Doris, Shaun Nichols, Stephen Stich and five anonymous reviewers.

[2] In each of the studies that follow, we found a statistically significant difference between intuitions about a morally good act and intuitions about a morally bad act, but one might well wonder how large each of those differences was. The answers are as follows. *Intentional action*: 33% vs. 82%. (All subsequent results are on a scale from 1 to 7.) *Deciding*: 2.7 vs. 4.6. *In favor*: 2.6 vs. 3.8. *In order to*: 3.0 vs. 4.6. *By*: 3.0 vs. 4.4. *Causation*: 2.8 vs. 6.2. *Doing/allowing*: 3.0 vs. 4.6.

[3] Surprisingly, there was also a significant gender x character interaction, whereby women tended to regard the act as more intentional when the agent had a bad character while men tended to regard the act as more intentional when the agent had a good character. I have no idea why this might be occurring, but it should be noted that this is just one of the many individual differences observed in these studies. Feltz and Cokely (2007) have shown that men show a greater moral asymmetry in intentional action intuitions when the vignettes are presented within-subject, and Buckwalter (2010) has shown that women show a greater moral asymmetry when they are asked about the agent's knowledge. Though not well-understood at the moment, these individual differences might hold the key to future insights into the moral asymmetries discussed here. (For further discussion, see Nichols & Ulatowski 2007.)

[4] Strikingly, recent research has shown that people's intuitions about intentional action can be affected by non-moral factors, such as judgments about the agent's own interests (Machery 2008; Nanay forthcoming), knowledge of conventional rules (Knobe 2007) and implicit attitudes (Inbar et al. 2009). This recent discovery offers us an interesting opportunity to test the present account. If we can come up with a general theory about how people's evaluations impact their thinking about alternative possibilities – a theory that explains not only the impact of moral judgments but also the impact of other factors – we should be able to generate predictions about the precise ways in which each of these other factors will impact people's intentional action intuitions. Such predictions can then be put to the test in subsequent experiments.

[5] There may be certain general theoretical reasons for adopting the view that people's representations of the agent's attitude have this continuous character, but the principal evidence in favor of it comes from the actual pattern of the experimental data. For example, suppose that instead of saying that the agent does not care at all about the bad side-effect, we say that the agent deeply regrets the side-effect but decides to go ahead anyway so as to achieve the goal. Studies show that people then tend to say that the side-effect was brought about *unintentionally* (Phelan & Sarkissian 2008; Sverdlik 2004). It is hard to see how one could explain this result on a model in which people have a unified way of thinking about all attitudes that involve the two features (1) foreseeing that an outcome will arise but (2) not specifically wanting it to arise. However, the result becomes easy to explain if we assume that people represent the agent's attitude, not in terms of sets of features (as I earlier believed; Knobe 2006), but in terms of a continuous dimension. We can then simply say that people take the regretful agent to be slightly more toward the 'con' side of the continuum and are therefore less inclined to regard his or her behavior as intentional.

[6] Very briefly, the suggestion is that intuitions in all three of these domains involve a capacity to compare reality to alternative possibilities. Thus, (a) intuitions about whether an agent acted freely depend on judgments about whether it was possible for her to choose otherwise, (b) intuitions about whether a person knows something depend on judgments about whether she has enough evidence to rule out relevant alternatives, and (c) intuitions about whether one event caused another depend on judgments about whether the second event would still have occurred if the first had not. Since moral judgments impact the way people decide which possibilities are relevant or irrelevant, moral judgments end up having an impact on people's intuitions in all three of these domains.

References

Adams, F. & Steadman, A. (2004a) Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis* 64:173-81.

Adams, F. & Steadman, A. (2004b) Intentional actions and moral considerations: Still pragmatic. *Analysis* 64:268-76.

Adams, F. and A. Steadman. 2007: Folk concepts, surveys, and intentional action. In C. Lumer (ed.). *Intentionality, deliberation, and autonomy: The action-theoretic basis of practical philosophy*. Aldershot: Ashgate Publishers.

Alicke, MD. (2000) Culpable control and the psychology of blame. *Psychological Bulletin* 126:556-74.

Alicke, MD. (2008) Blaming badly. *Journal of Cognition and Culture* 8:179-186.

Beebe, J. R. & Buckwalter W. (forthcoming) The epistemic side-effect effect. *Mind & Language*.

Buckwalter, W. (2010) Gender and epistemic intuition. Unpublished manuscript. City University of New York.

Byrne, R. (2005) *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT Press.

Carey, S. & Spelke, E. (1996) Science and core knowledge. *Philosophy of Science*, 63:515-533

Chapman, L. & Chapman, J. (1967) Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*. 72: 193-204.

Churchland, P. (1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78(2):67-90.

Cushman, F. (2010) Judgments of morality, causation and intention: Assessing the Connections. Unpublished manuscript, Harvard University.

Cushman, F. & Mele, A. (2008) Intentional action: Two-and-a-half folk concepts? In: *Experimental Philosophy*, ed. J. Knobe & S. Nichols. Oxford University Press.

Cushman, F., Knobe, J. & Sinnott-Armstrong, W. (2008) Moral appraisals affect doing/allowing judgments. *Cognition* 108:353-80.

Damasio, A.R., Tranel, D. & Damasio, H. (1990) Individuals with socio-pathic behavior caused by frontal damage fail to respond autonomically to social stimuli. *Behavioural Brain Research* 41:81–94.

De Villiers, J., Stainton, R. & Szatmari, P. (2006) Pragmatic abilities in autism spectrum disorder: A case study in philosophy and the empirical. *Midwest Studies in Philosophy* 31:292-317.

Ditto, P., Pizarro, D. & Tannenbaum, D. (2009) Motivated moral reasoning. In: *Moral judgment and decision making: The psychology of learning and motivation*, ed. D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin. Elsevier.

Driver, J. (2008a) Attributions of causation and moral responsibility. In: *Moral psychology volume 2: The cognitive science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong. MIT Press.

Driver, J. (2008b) Kinds of norms and legal causation: Reply to Knobe and Fraser and Deigh. In: *Moral psychology volume 2: The cognitive science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong. MIT Press.

Feltz, A. & Cokely, E. (2007) An anomaly in intentional action ascription: More evidence of folk diversity. *Proceedings of the Cognitive Science Society*.

Goldman, A. (1970) *A theory of human action*. Prentice-Hall, Inc.

Goldman, A. (2006) *Simulating minds: The philosophy, psychology and neuroscience of mindreading*. Oxford University Press.

Gopnik, A. & Meltzoff, A. (1997). *Words, thoughts and theories*. Cambridge: MIT Press.

Gopnik, A. & Wellman, H. (1992). Why the child's theory of mind really is a theory. *Mind & Language* 7:145-71.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T. & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111: 1-31.

Grice, H. P. (1989) *Studies in the way of words*. Harvard University Press.

Guglielmo, S. & Malle, B.F. (2009a) Can unintended side-effects be intentional? Solving a puzzle in people's judgments of intentionality and morality. Unpublished manuscript, Brown University.

Guglielmo, S. & Malle, B.F. (2009b) The timing of blame and intentionality: testing the moral bias hypothesis. Unpublished manuscript, Brown University.

Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108:814-34.

Hindriks, F.A. (forthcoming) Intentional action and the praise-blame asymmetry. *Philosophical Quarterly*.

Hitchcock, C. & Knobe, J. (forthcoming) Cause and norm. *Journal of Philosophy*.

Inbar, Y., Pizarro, D.A., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion* 9*:*435-439.

Kahneman, D., & Miller, D. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review* 93:136-153.

Kelley, H. H. (1967) Attribution theory in social psychology. In: *Nebraska Symposium on Motivation*, ed. D. Levine. University of Nebraska Press.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. & Damasio, A. (2007) Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446:908-11.

Knobe, J. (forthcoming) Action tree and moral judgment. *Topics in Cognitive Science*.

Knobe, J. (2003a) Intentional action and side effects in ordinary language. *Analysis* 63:190-3.

Knobe, Joshua. (2004) Intention, intentional action and moral considerations. *Analysis* 64:181-187.

Knobe, J. (2003b) Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology* 16:309-24.

Knobe, J. (2004a) Intention, intentional action and moral considerations. *Analysis* 64:181-7.

Knobe, J. (2004b) Folk psychology and folk morality: Response to critics. *Journal of Theoretical and Philosophical Psychology* 24(2):270-79.

Knobe, J. (2007) Reason explanation in folk psychology. *Midwest Studies in Philosophy* 31:90-107.

Knobe, J. & Burra, A. (2006) Intention and intentional action: A cross-cultural study. *Journal of Culture and Cognition* 6:113-32.

Knobe, J. & Fraser, B. (2008) Causal judgment and moral judgment: Two experiments. In: *Moral psychology volume 2: The cognitive science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong. MIT Press.

Knobe, J. & Roedder, E. (forthcoming) The ordinary concept of valuing. *Philosophical Issues*.

Kunda, Z. (1990) The case for motivated reasoning. *Psychological Bulletin* 108(3):480-498.

Leslie, A., Knobe, J. & Cohen, A. (2006) Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment. *Psychological Science* 17:421-7.

Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language* 23:165-189.

Martin, K. (2009) An experimental approach to the normativity of 'natural'. Paper presented at the Annual Meeting of the South Carolina Society for Philosophy, Rock Hill, South Carolina, 2009.

McArthur, L & Post, D. (1977). Figural emphasis and person perception. *Journal of Experimental Social Psychology* 13: 520-535.

McCann, H. (2005) Intentional action and intending: Recent empirical studies. *Philosophical Psychology* 18:737-48.

McCloy, R. & Byrne, R. (2000). Counterfactual thinking about controllable events. *Memory and Cognition* 28: 1071-1078.

Mikhail, J. (2007) Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences* 11:143-52.

Murphy, G. L. & Medin D. L. (1985) The roles of theories in conceptual coherence. *Psychological Review* 92:289-316.

N'gbala, A., & Branscombe, N.R. (1995). Mental simulation and causal attribution: When simulating an event does not affect fault assignment. *Journal of Experimental Social Psychology* 31:139-162.

Nadelhoffer, T. (2005) Skill, luck, control, and folk ascriptions of intentional action. *Philosophical Psychology* 18:343-54.

Nadelhoffer, T. (2006a) Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality. *Philosophical Explorations* 9:203-20.

Nadelhoffer, T. (2006b) On trying to save the Simple View. *Mind & Language* 21:565-586.

Nanay, B. (forthcoming). Morality of modality? What does the attribution of intentionality depend on? *Canadian Journal of Philosophy.*

Nichols, S. & Ulatowski, J. (2007) Intuitions and individual differences: The Knobe effect revisited. *Mind & Language* 22:346-65.

Nyholm, S. (2009) Moral judgments and happiness. Unpublished manuscript, University of Michigan.

Pettit, D. & Knobe, J. (forthcoming) The pervasive impact of moral judgment. *Mind & Language*.

Phelan, M. & Sarkissian, H. (2008) The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies* 138: 291-298.

Phillips, J. & Knobe, J. (2009) Moral judgments and intuitions about freedom. *Psychological Inquiry* 20:30-6.

Premack, D., & Woodruff, G. (1978c). Does the chimpanzee have a theory of mind. *Behavioral and Brain Sciences* 1:515-526.

Roese, N. (1997). Counterfactual thinking. *Psychological Bulletin* 121:133-148.

Roxborough, C. & Cumby, J. (2009) Folk psychological concepts: Causation. *Philosophical Psychology* 22:205-13.

Shepard, J. (2009) The side-effect effect in Knobe's environment case and the Simple View of intentionality. Unpublished Manuscript, Georgia State University.

Sloman, S. (2005) Causal models: How people think about the world and its alternatives. Oxford University Press.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology* 4: 165–173.

Solan, L. & Darley, J. (2001) Causation, contribution, and legal liability: An empirical study. *Law and Contemporary Problems* 64:265-98.

Surian, L., Baron-Cohen, S., & van der Lely, H. K. J. (1996) Are children with autism deaf to Gricean maxims? *Cognitive Neuropsychiatry* 1:55-71.

Sverdlik, S. (2004) Intentionality and moral judgments in Commonsense Thought about Action. *Journal of Theoretical and Philosophical Psychology* 24:224-236.

Tannenbaum, D., Ditto, P. & Pizarro, D. (2009) Different moral values produce different judgments of intentional action. Unpublished Manuscript, University of California, Irvine.

Tetlock, P.E. (2002). Social-functionalist frameworks for judgment and choice: The intuitive politician, theologian, and prosecutor. *Psychological Review* 109: 451-472.

Turner, J. (2004) Folk intuitions, asymmetry, and intentional side effects. *Journal of Theoretical and Philosophical Psychology* 24:214-9.

Ulatowski, J. (2009) Action under a description. Unpublished manuscript, University of Wyoming.

Woodward, J. (2004). *Making things happen: A theory of causal explanation.* Oxford: Oxford University Press.

Wright, J.C., & Bengson, J. (2009) Asymmetries in judgments of responsibility and intentional action. *Mind & Language* 24(1):24-50.

Young, L., Cushman, F., Adolphs, R., Tranel, D. & Hauser, M. (2006) Does emotion mediate the effect of an action's moral status on its intentional status? Neuropsychological evidence. Journal of Cognition and Culture 6:291-304.

Zalla, T., Machery, E. & Leboyer, M. (2010) Intentional action and moral judgment in Asperger Syndrome and high-functioning autism. Unpublished Manuscript, Institut Jean-Nicod.