BEHAVIORAL AND BRAIN SCIENCES (2010) **33**, 315–365
doi:10.1017/S0140525X10000907

Person as scientist, person as moralist

Joshua Knobe
Program in Cognitive Science and Department of Philosophy, Yale University, New Haven, CT 06520-8306
joshua.knobe@yale.edu
http://pantheon.yale.edu/~jk762/

Abstract: It has often been suggested that people's ordinary capacities for understanding the world 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 actually influence the intuitions they hold both in folk psychology and in causal cognition. The present target article distinguishes two basic approaches to explaining such effects. One approach would be to say that the relevant competencies are entirely non-moral but that some additional factor (conversational pragmatics, performance error, etc.) then interferes and allows people's moral judgments to affect their intuitions. Another approach would be to say that moral considerations truly do figure in workings of the competencies themselves. I argue that the data available now favor the second of these approaches over the first.

Keywords: Causal cognition; moral cognition; theory of mind

1. Introduction

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, and

JOSHUA KNOBE is Assistant Professor of Cognitive Science and Philosophy at Yale University. He is one of the founding members of the "experimental philosophy" movement.

 0140-525X/10 $40.00

the like, 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.

2. 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.

2.1. 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 seems 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.

2.2. 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

https://doi.org/10.1017/S0140525X10001810

place." This would certainly be a reasonable approach to the study of religious discourse, but the key claim of the person-as-scientist theory 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:

1. 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.
2. 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.
3. 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 particular differences between these different 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 strategy 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 I shall discuss it in further detail later. 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.

3. 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.

3.1. 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 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.[1] 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., 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 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 4-year-old children (Leslie et al. 2006a). 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.

3.2. Further mental states

As it happens, though, the impact of moral considerations does not appear to be limited to people's use of the word "intentionally." The very same effect also arises for numerous other expressions: "intention," "deciding," "desire," "in favor of," "advocating," and so forth.

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, are 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 2004a). 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, and they tend to disagree with the claim that the agent decided to help (Pettit & Knobe 2009).

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 2004b; 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 in numerous other domains. 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 2009). And similar effects have been observed for people's use of many other concepts: *desiring*, *intending*, *choosing*, and so forth (Pettit & Knobe 2009; 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

https://doi.org/10.1017/S0140525X10001810

a whole range of different concepts used to pick out mental states and processes.

3.3. 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 with the tree in Figure 1.

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, for example, 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 2004b; forthcoming). Hence, people are willing to say:

> The chairman harmed the environment in order to increase profits.

but not:

> The chairman helped the environment in order to increase profits.

And, similarly, they are willing to say:

> The chairman increased profits by harming the environment.

but not:

> 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 (see Fig. 2).

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).

3.4. 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 2009; 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

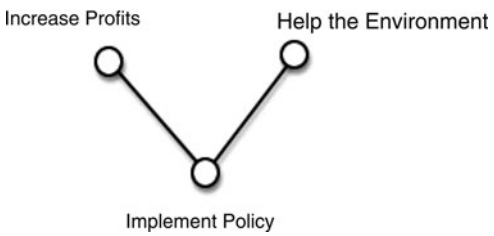

Figure 1. Action tree for the help case.

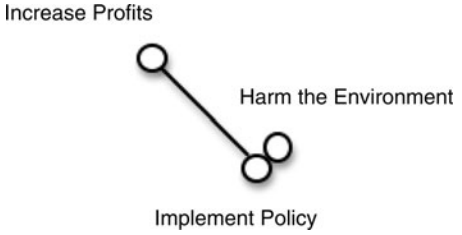

Figure 2. Action tree for the harm case.

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 et al. 2008). All in all, the evidence seems strongly to suggest that people's moral judgments are actually impacting their causal intuitions.

3.5. 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 someone 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* versus *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.

3.6. 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 2009), *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.

4. 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 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 relevant 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

https://doi.org/10.1017/S0140525X10001810

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.

4.1. 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, et cetera – 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 of this relationship looks something like what's shown in Figure 3

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

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

it be that their moral judgments affect their intuitions about causation?

To resolve this question, one might develop a model that goes more like the one shown in Figure 4

In this revised model, there is a reciprocal relationship between people's blame judgments and their intuitions about intention, causation, et cetera. 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 (Ditto et al. 2009; Haidt 2001). 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 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,

Figure 4. Motivational bias account of blame ascription.

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 2008; Machery 2008; McCann 2005; Nichols & Ulatowski 2007; Turner 2004; Wright & Bengson 2009; Young et al. 2006).

4.1.1. Neuropsychological studies. 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 et al. (2006) did just that. They took the cases of the corporate executive who harms or helps the environment 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 situations 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 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.

4.1.2. Types of moral judgment. 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, and the like.

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 proceeds in a quite different way (see Fig. 5).

This model involves a quite radical rejection of the view that people's intuitions about intention, causation, et cetera, 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 (in press) 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 (2009) 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

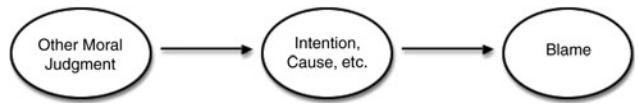

Figure 5. Distinct processes of moral judgment.

https://doi.org/10.1017/S0140525X10001810

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.[2]

In the face of these new results, friends of the motivational bias view might simply 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."

4.1.3. Cause and blame. 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 the data don't actually suggest that people's causal intuitions are being influenced by a judgment that the agent is to blame for the outcome. Instead, the data appear to suggest that these intuitions are 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:

(a) The professor is to blame for the outcome (the receptionist's lack of pens).

(b) 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 the 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 2009). 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.

4.1.4. Conclusion. Of course, it might ultimately prove possible to wriggle out of all of these difficulties and show that the data reviewed here do 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; Ditto et al. 2009; Haidt 2001), but when it comes to the specific effects under discussion here, there are no such experiments. Instead, the argument always proceeds by drawing on experimental studies in one domain to provide evidence about the psychological processes at work in another (see, e.g., Nadelhoffer 2006a). That is, the argument has roughly the form: "This explanation turned out to be true for so many other effects, so it is probably true for these ones, 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.

4.2. 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 the corresponding 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 nonfunctional 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 actual 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 concept of intentional action 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 of causation 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 further, 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 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.

4.2.1. Patient studies. 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 did exactly that in a recent study (Zalla et al. 2010). 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 et al. 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 neurotypical 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 observed in previous studies. Just like neurotypical subjects, people with Asperger's tended to say that the chairman harmed the environment intentionally but helped it unintentionally. This result suggests that the pattern displayed by subjects in earlier studies is not, in fact, a product of their mastery of complex pragmatic principles.

4.2.2. Cancelation. Of course, the study of linguistic deficits in people with Asperger's brings up a host of complex

https://doi.org/10.1017/S0140525X10001810

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 2004b; Nichols & Ulatowski 2007; for a review, see Nadelhoffer 2006c).

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 these 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.

4.2.3. Other effects. 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 actual 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.

4.3. 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 about the nature of the basic competencies people use to understand their world.

5. 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.

5.1. 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 from 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, for example, to consider the possibility that the car could have been levitating in the air, and then think: "Notice that this 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 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. Because 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.[3]

5.2. 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

https://doi.org/10.1017/S0140525X10001810

people's way of thinking about alternative possibilities. This would be a very complex task, and I 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. So we will be assuming that people can represent the agent as strongly opposed, as strongly in favor, 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*. [4] (See Fig. 6.)

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. (See Fig. 7.)

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 the agent's 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.

Figure 6. Continuum of attitude ascription.

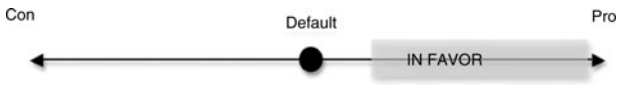

Figure 7. Criteria for ascription of "in favor."

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. (See Fig. 8.)

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 or her attitude goes sufficiently far beyond this default (Fig. 9).

In this new representation, the agent's actual attitude remains at exactly the same point it was above (in Fig. 8), but its position relative to the default is now quite different. This time, the attitude 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 the agent was in favor of harming – which, in fact, is exactly what the experimental results show.

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

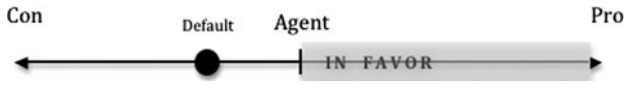

Figure 9. Representation of the continuum for the harm case.

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 they can use to make sense of the behavior they actually observe.

5.3. 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[5] (Hitchcock & Knobe 2009; Pettit & Knobe, 2009; 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.

6. Conclusion

This target article 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.

ACKNOWLEDGMENTS

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

NOTES

1. 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.

2. 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.)

3. 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 2010), 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.

4. 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.

5. 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. Because 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.

Open Peer Commentary

Competence: What's in? What's out? Who knows?

doi:10.1017/S0140525X10001652

Joshua Alexander,[a] Ronald Mallon,[b] and Jonathan M. Weinberg[c]

[a]*Philosophy Department, Siena College, Loudonville, NY 12211;* [b]*Department of Philosophy, University of Utah, Salt Lake City, UT 84112;* [c]*Department of Philosophy, Indiana University, Bloomington, IN 47405-7005.*

jalexander@siena.edu **http://www.siena.edu/pages/1855.asp**
rmallon@philosophy.utah.edu
http://www.philosophy.utah.edu/faculty/mallon/
jmweinbe@indiana.edu
http://www.indiana.edu/~phil/Faculty/Individual%20Pages/Weinberg.html

Abstract: Knobe's argument rests on a way of distinguishing performance errors from the competencies that delimit our cognitive architecture. We argue that other sorts of evidence than those that he appeals to are needed to illuminate the boundaries of our folk capacities in ways that would support his conclusions.

Joshua Knobe argues that the various moral inflections of our folk psychology are part of "the competencies people use to understand the world" (target article, sect. 4, para. 1), a hypothesis that he contrasts with the claim that "certain additional factors are somehow 'biasing' or 'distorting' people's cognitive processes and thereby allowing their intuitions to be affected by moral judgments" (sect. 4, para. 1). However, Knobe really never makes clear exactly what makes something "inside" or "outside" a competence. Clearly, both he and his past interlocutors have taken motivated cognition and pragmatic factors to count as "interfering," rather than as part of the competence. But what can ground such judgments? We worry that any non-stipulative way of answering this question that plausibly excludes motivation and pragmatic considerations, can also be used to insist that moral considerations are "outside of" or "external to" the competence under consideration. We are not disputing the empirical facts that he does muster; rather, our concern is with a further theoretical interpretation he wants to place on those facts, which we argue is unwarranted.

One natural way to circumscribe the boundaries of a competence is *bottom-up*, by appealing to a fairly literal, physical notion of containment provided by neuroanatomy. But neither Knobe nor his interlocutors muster any such neuroanatomical evidence, so this sort of approach is not a good contender.

A more promising way of approaching questions of competence is to begin with a high-level characterization of the function that a cognitive process is supposed to compute, and on this basis attempt to specify an algorithm for computing that function and to address questions of actual physical implementation (see Mallon 2007). Once we are clear about what task a cognitive process is supposed to execute, constraints or problems in the execution of the task can be identified. According to this *top-down* approach to competence, then, what allows us to describe something as interfering with a cognitive process is a substantive account of the work the process is supposed to be doing.

The problem with taking this kind of approach here is that there isn't a settled account of what sort of job our folk psychological judgments are supposed to do. What we have are two different models, each of which stipulates what function is supposedly being calculated by our folk psychological judgments, and thus, what is and is not part of our competence with such judgments. On Knobe's model, pragmatics and motivation may indeed properly lie outside the competence. But on his opponent's model, the exact same line of reasoning would apply to the source of the morality effects that Knobe is appealing to. There is no "supposed to" to be found within those sorts of findings, and so where to draw an inside/outside line is, thus far, an empirical free move – stipulated, not discovered.

One can contrast the situation regarding our folk-psychological capacities with the comparatively much better established taxonomy of competences in both language and vision. For example, linguists are used to separating out the semantic, syntactic, phonological, and pragmatic components of our overall linguistic capacities. This division has proved empirically fruitful, and it is grounded in bottom-up considerations as well (such as deficit patterns due to various lesions). In debates about language, then, it makes sense that showing a phenomenon to be a proper part of one member of that partition is thereby a good reason to reject it as part of some other member. Interestingly, we see exactly this dynamic in earlier stages of the debate about the side-effect effect, which for several years was explicitly framed in terms of whether there was a moral dimension to the semantic component of our "intentionally" discourse. And so it made sense, in the context of the debate so construed, to take a pragmatic explanation of the side-effect effect to preempt an explanation of it in semantic terms. The existing framework legitimated ruling pragmatics to be "outside" of semantics. But in this

target article, Knobe switches from a debate about semantic competence to a debate about competence in some abstract sense. And we think Knobe has made the right decision to move away from the semantic debate, in part because of some Quinean pessimism about certain forms of conceptual analysis (see Alexander et al. 2010). But more importantly, Knobe's own favored account (sect. 5) locates the source of this moral inflection in a mechanism for the allocation of cognitive resources, selecting different alternatives for cognitive attention. He writes, "The basic idea is just that people's intuitions ... rely on a comparison between the actual world and certain alternative possibilities. Because people's moral judgments influence the selection of alternative possibilities, these moral judgments end up having a pervasive impact" (sect. 5.1, para. 8). But this influence on cognitive attention would count as part of general cognition, but not part of semantics, in the traditional linguistic taxonomy, and thus would have counted as an "external" causal factor in the old debate.

Given his current choice of hypothesis, this shift away from a specifically semantic framing of the issues to a more generic one makes sense. Unfortunately, in doing so, he has abandoned one set of resources for underwriting an inside/outside distinction without replacing them with something else. The challenge he faces is how to substantiate such a distinction in a way that both (1) isn't merely stipulative and (2) puts pragmatics and motivation on the "outside" and cognitive attention on the "inside."

We think that, in order to do so, other sorts of evidence will be needed than those that Knobe appeals to in his article. Perhaps an evolutionary or teleological argument could ground a top-down approach here; or perhaps neuroanatomical evidence could ground a bottom-up approach; or perhaps – and where we would place our bets – further re-refinement of the basic question is still in order.

Culpable control or moral concepts?

doi:10.1017/S0140525X10001664

Mark Alicke[a] and David Rose[b]

[a]*Department of Psychology, Ohio University, Athens, OH 45701;* [b]*Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15235.*

alicke@ohio.edu

davidros@andrew.cmu.edu

Abstract: Knobe argues in his target article that asymmetries in intentionality judgments can be explained by the view that concepts such as intentionality are suffused with moral considerations. We believe that the "culpable control" model of blame can account both for Knobe's side effect findings and for findings that do not involve side effects.

Virtually everyone, including both professional and lay evaluators of human behavior, agrees that to praise or blame an agent, the agent must have acted intentionally, with foresight of the consequences, and must have caused the outcome. In a perfect evaluative world, assessments of intentionality, foresight, and causation would be made independently of the judge's expectations, affective and attitudinal reactions, and moral beliefs and predilections. By now, however, there is abundant evidence that such factors have a powerful and pervasive influence on intentionality, foresight, and causation judgments. In his target article, Knobe offers an alternative to the view that such influences are motivated by the desire to justify praising or blaming an agent who evokes positive or negative evaluative reactions.

Knobe has argued that the reason that concepts such as intentionality, causation, and foresight are influenced by moral considerations is because these concepts are suffused with moral considerations (i.e., moral considerations figure into the underlying competencies). To illustrate his position, he uses his well-traveled example of an executive who knows that, as a side effect of initiating a certain program, the environment will be helped or harmed, but whose only concern is to increase profits. Knobe assumes that in the "help" scenario, the normative (or "default") expectation would be to have at least a moderately *pro*-attitude toward helping the environment. The default expectation in the "harm" scenario would be for a moderately anti-attitude (or in Knobe's term, *con*-attitude) toward harming environment. In the "help" case, the executive doesn't meet the threshold required for having a *pro*-attitude; whereas in the "harm" case, the executive's failure to endorse a *con*-attitude places him in the range of plausibly having a *pro*-attitude. Thus, in the latter case but not the former, the executive is thought to have acted intentionally because he apparently has a *pro*-attitude toward the harmful outcome. Ascriptions of intentionality are thus dependent on the sort of attitude an evaluator thinks an agent should have about a particular outcome. Intentionality (and presumably other concepts such as causation and foresight) is applied when the agent's presumed attitude crosses the evaluator's normative threshold.

Our alternative to Knobe's position – the *culpable control* model of blame (Alicke 1992; 2000; Alicke & Zell 2009; Alicke et al. 2010) – can explain the environmental harming/helping findings without positing that concepts such as intentionality are inherently moral. Further, it can explain cases other than the specialized side-effect scenarios that are the focus of Knobe's theory. The culpable control model assumes that positive and negative evaluative reactions – which are judgments of right or wrong, good or bad, or approval or disapproval – to the people involved in an event, their actions, and the outcomes that ensue can induce social perceivers to process information in a "blame validation" mode. Blame validation involves interpreting the evidence regarding intentionality, causation, and foresight in a way that justifies praising an agent who elicits positive evaluations or blaming one who arouses negative evaluations.

The culpable control model explains Knobe's findings by assuming that social perceivers view the environment-harming executive as a major jerk (i.e., one who arouses strong negative evaluations), but view the environment-helping executive as only a minor one. Imputing intentionality to the environment-harming executive, therefore, validates social perceivers' negative evaluative reactions, and in turn, supports a blame attribution. The culpable control model, therefore, does not require the assumption that concepts such as intentionality, causation, and foresight are suffused with moral considerations. Rather, the influences of these evaluations can be explained in terms of the desire to blame an agent whose actions arouse strong disapproval.

It is important to note that most of Knobe's examples apply to cases where foreseen but unintended side-effects occur, which narrows the application of the theory. Furthermore, Knobe's example is not an optimal one for considering the relative merits of his position and the culpable control model because his default assumption regarding an agent's attitude is confounded with positive and negative evaluations of the agent's goals and actions and the outcomes that occur. For example, in the "harm" scenario, the agent's indifference toward harming the environment diverges from the attitude we would expect most agents to have in this situation; but this indifference, as well as both the agent's decision to let the environment be harmed and the fact that the environment ultimately is harmed, also provides a basis for negative evaluative reactions. So, we need some other case to differentiate these two views.

In an early set of studies (Alicke 1992, Study 1), participants learned that a young man was speeding home to hide either an anniversary present or a vial of cocaine from his parents before they arrived home. A car accident occurred under somewhat ambiguous circumstances: It could have been due in part to his speeding, but also to environmental impediments such as a partly obscured stop sign. The study's results were clear: When the driver's motive was undesirable (i.e., to hide cocaine), his driving was cited as far more causal than the environmental obstacles. However, precisely the opposite was true when his

motive was to hide an anniversary present. This case would be very difficult for Knobe's theory to explain: Why would anyone assume that the driver who was speeding home to hide cocaine had a *pro*-attitude towards causing car accidents, and that the driver who was speeding to hide an anniversary present did not? The more plausible alternative, based on the culpable control model, is that negative evaluations of the driver whose motive was to hide cocaine induced participants to skew the evidential criteria for causation to support their desire to blame him.

In sum: Knobe's position is plausible in cases involving foreseen but unintended side-effects, but it has some trouble explaining cases outside of this narrow scope. The culpable control model can explain many cases involving side-effects, as well as most cases that do not involve side-effects, and can do so without claiming that concepts such as intentionality, causation, and foresight are suffused with moral considerations.

Person as moralist and scientist

doi:10.1017/S0140525X10001676

Marcus Vinícius C. Baldo[a] and Anouk Barberousse[b]

[a] *"Roberto Vieira" Laboratory of Sensory Physiology, Department of Physiology and Biophysics, Institute of Biomedical Sciences, University of São Paulo, SP 05508-900, São Paulo, Brazil;* [b] *Institut d'Histoire et de Philosophie des Sciences et des Techniques (IHPST), UMR 8590, CNRS, Université Paris 1, ENS, 75006 Paris, France.*

baldo@usp.br

http://www.fisio.icb.usp.br/~vinicius/barberou@heraclite.ens.fr

http://www-ihpst.univ-paris1.fr/en/4,anouk_barberousse.html

Abstract: Scientific inquiry possibly shares with people's ordinary understanding the same evolutionary determinants, and affect-laden intuitions that shape moral judgments also play a decisive role in decision-making, planning, and scientific reasoning. Therefore, if ordinary understanding does differ from scientific inquiry, the reason does not reside in the fact that the former (but not the latter) is endowed with moral considerations.

According to Knobe's central thesis, we are "moralizing creatures," with moral judgments lying at the core of the competencies we use to make sense of our actions and ourselves. By "ordinary understanding," Knobe means the way people make sense of the world without having any scientific education. He argues that human cognition, in general, is intrinsically and inescapably moral, in the sense that people just do not make sense of certain situations without performing proto-moral judgments at the very time they perceive them. However, the target article's argument is not as generally applicable as Knobe claims: the article does not address people's cognition in general, but only how we, as human beings, perceive and interpret human interactions, as is made clear in the reported experiments. While this topic in fact belongs to the study of causal cognition, it is far from exhausting it.

Also, the way Knobe contrasts his main thesis with other claims that have been made before seems to be misleading, promoting an erroneous interpretation of his own thesis. He opposes his view about cognition to the idea that the functioning of the human mind mirrors the functioning of scientists (Gopnik 1996; Gopnik & Schulz 2004; Gopnik & Tenenbaum 2007). Despite Knobe's insistence on criticizing this view, it is the wrong opponent to his own thesis. What he is actually attacking is a view according to which humans' perception of humans' interactions is objective, in the sense of being devoid of any moral commitment. Gopnik's view, for instance, is about the development of cognition as much as about cognition in adults, independently of their being impregnated or not with moral considerations.

Actually, if we adopt the thesis according to which scientific inquiry is only a very refined manifestation of our effort to make sense of the world, sharing with people's ordinary understanding the same evolutionary roots, then Knobe's claim that ordinary understanding does not build itself in the same manner as scientific understanding does, loses much of its strength. The justification of this thesis will follow in two steps.

An increasing body of evidence, from fields such as psychology, anthropology, and neuroscience, lends support to the idea that quick and automatic affect-laden intuitions indeed shape higher levels of human reasoning, which belong to a slower and phylogenetically newer set of cognitive resources (Eslinger & Damasio 1985; Fiske 1992; Greene & Haidt 2002; Haidt 2001; 2007; Moll et al. 2005; Prinz 2006). Thus, the first step is to consider emotional and affective factors as the driving forces behind moral judgments. This view is by no means new; it goes back at least to David Hume's proposal that moral reasoning is driven by moral emotions and intuitions: "Reason is, and ought only to be the slave of the passions" (Hume 1739/2000).

The second step is to dissolve the concept of "moral judgment" by no longer envisaging it as a single entity, but rather as a compound cognitive act. Based on neuroscientific evidence, it is becoming increasingly clear that there are no specific regions of the brain responsible for moral judgments, which would be the combined result of basic processes involving abstract reasoning, its emotional content, and possibly other cognitive factors (Greene & Haidt 2002). Lending support to this view, the "affect as information" hypothesis, from social psychology, emphasizes the importance of people's mood and feelings when making decisions and judgments (Haidt 2001). In the same vein, Damasio's "somatic marker" hypothesis points to the role of emotional experiences in guiding decision-making by ascribing affective valence to behavioral options, and this has been substantiated by empirical data originating from several clinical and neuroimaging studies (Eslinger & Damasio 1985; Bechara et al. 2000).

In conclusion, Knobe's article has the merit of reinforcing, with both experimental facts and theoretical considerations, the not often recognized importance of moral values in assembling apparently neutral and objective evaluations. However, we believe that such an idea must be brought to a wider scenario, in which we could integrate basic neurophysiological mechanisms underlying emotional states and decisional processes into a framework able to account for more elaborate cognitive tasks, such as planning, moral judgments, and scientific reasoning. Finally, granting to Knobe that we inescapably moralize our perception of human interactions, it would be interesting to say a word about the relationships between this intuitive grasping and the scientific grasping that psychologists and sociologists try to achieve. What does happen in a psychologist's mind when she studies someone looking at two people interacting? Such a question is likely to allow one to go deeper into the implications of Knobe's proposition.

ACKNOWLEDGMENT

This work was partially supported by Brazilian funding agencies (FAPESP/CNPq).

Reasoning asymmetries do not invalidate theory-theory

doi:10.1017/S0140525X10001688

Karen Bartsch and Tess N. Young

Psychology Department 3415, University of Wyoming, Laramie, WY 82071-3415.

bartsch@uwyo.edu

tyoung14@uwyo.edu

http://uwadmnweb.uwyo.edu/psychology/displayfaculty.asp?facultyid=1285

Abstract: In this commentary we suggest that asymmetries in reasoning associated with moral judgment do not necessarily invalidate a theory-

theory account of naïve psychological reasoning. The asymmetries may reflect a core knowledge assumption that human nature is prosocial, an assumption that heightens vigilance for antisocial dispositions, which in turn leads to differing assumptions about what is the presumed topic of conversation.

We question Knobe's thesis that in acknowledging the asymmetries observed in reasoning associated with moral judgment we must perforce abandon a "theory-theory" characterization of naïve psychological reasoning. Certainly there are pervasive asymmetries, at least when the situation affords opportunity for moral evaluation, but such asymmetries may not, in themselves, invalidate a theory-theory perspective.

We suspect that underlying the asymmetries is a tendency to view prosocial dispositions as the norm, a tendency that makes us especially vigilant about antisocial dispositions. In other words, we firmly expect others to have prosocial intentions and behavior. When asked about actions that involve harm-infliction or rule breaking, we assume these anomalies are the events that require explanation and attention. This vigilance, which recognizes moral culpability as a conspicuous and interesting phenomenon, can account for the empirical findings in question without compelling rejection of a theory-theory characterization of naïve psychological reasoning. That is, just because we bring to our reasoning something like a base-rate assumption, or perhaps even a "core knowledge" assumption (à la Spelke & Kinzler 2007) concerning prosocial intentions and behavior, it does not follow that we are essentially unscientific or irrational in the reasoning that ensues.

Even if we begin with a default assumption about its nature, we may nevertheless view human behavior through a theory-like framework that can be rationally revised through experience (e.g., Gopnik & Wellman 1992). In essence, we suggest that, despite the arguments raised by Knobe against motivational and conversational pragmatics hypotheses, there is some such alternative view that can account for the empirical findings without invalidating a theory-theory characterization of naïve psychological reasoning.

Knobe's point regarding the pervasive nature of observed asymmetries in psychological reasoning, at least in reasoning that invites moral judgment, is well supported. But the asymmetries may simply reveal that people are poised to seize on moral failings and treat them as focal. They are less inclined to view prosocial activities as warranting explanation and evaluation. In keeping with the widely popular notion that we have a deep-seated intuitive sensitivity regarding harm (e.g., Haidt 2001; Hoffman 2000; Turiel 2006), perhaps we exhibit what a theory-theorist might consider a core knowledge bias when reasoning about others. Specifically, we view good intentions and behavior as the norm and are consequently hypersensitive to deviations from it. When asked about people who act badly, we assume the deviation is what we should attend to and explain (in a Gricean fashion), and respond accordingly.

On this view, when asked whether the chairman of the board has intentionally *harmed* the environment by instituting a new program that he knows would create harm in addition to increasing profits, people respond affirmatively. Such a response, we think, reflects vigilant attention to someone *knowingly* acting in a way that causes harm. The same interpretation applies to the findings reflecting variations on the terms ("deciding," "desire," "advocating," etc.). In these cases, the unacceptable, yet fully cognizant, behavior is simply assumed to be the topic of conversation. But when asked whether the chairman intentionally *helped*, rather than harmed, the environment, people respond differently – because helping is not a weird and worrying thing that commands vigilant attention. A question about whether helping was intended is interpreted as a query about whether the chairman had that outcome as a *goal*. In a world where prosocial aims are assumed to be the norm, we are not inclined to give the chairman moral credit for the outcome because the original goal, according to the narrative, was to increase profit. These asymmetrical responses do not reflect irrational reasoning; they reflect different assessments of what is the focal issue, or curiosity, to be addressed and explained. It may be that such asymmetries are universal because there is a core knowledge presupposition about the prosocial nature of people, possibly from very early in development (see Hamlin et al. 2007). This assumption, in itself, does not mean the reasoning is illogical or unscientific, but it does influence the topic presumed to warrant explanation and discussion.

How does this interpretation account for people's judgments of causation? When both an administrative assistant and a faculty member are said to have removed pens from the receptionist's desk, the former with permission and the latter illicitly, people indicate that the professor caused the problem. This response may also reflect vigilant attention to the anomalous bad behavior, the event that requires explanation, and a corresponding shift in conversational focus. Knobe astutely notes that it is not just that the behavior is strange or unusual (see, e.g., Roxborough & Cumby 2009) but specifically that the behavior is morally bereft. So it seems it is moral culpability in general that is regarded as anomalous, not the specific actions of a certain person or even a group of people. Moreover, it does not matter whether in fact the final outcome is actually good (Hitchcock & Knobe 2009, cited in the target article); the focus continues to be on the "bad" behavior. So – and here we think we agree with Knobe – it is not accurate to characterize what we bring to these reasoning tasks as purely a base-rate assumption regarding any specific action; rather, we simply assume that people will usually have benign intentions and act in a prosocial fashion.

In one respect, our interpretation resembles Knobe's. We agree that moral considerations are being taken into account in the reasoning reported in the literature. And it may be objected that our interpretation is simply another version of the motivational and conversational pragmatics hypotheses refuted by Knobe. But we think it is different to view people as bringing to their reasoning core content assumptions about human nature, assumptions that influence the presumed focus of conversations. The fact that people have expectations about others' prosocial and antisocial proclivities does not in itself mean that reasoning about causality in the social realm is irrational, impervious to experience, or otherwise at odds with scientific theorizing.

"Stupid people deserve what they get": The effects of personality assessment on judgments of intentional action

doi:10.1017/S0140525X1000169X

Berit Brogaard

Department of Philosophy and Department of Psychology, University of Missouri–St. Louis, St. Louis, MO 63121-4400.

brogaardb@umsl.edu **http://sites.google.com/site/brogaardb/**

Abstract: Knobe argues that people's judgments of the moral status of a side-effect of action influence their assessment of whether the side-effect is intentional. I tested this hypothesis using vignettes akin to Knobe's but involving economically or eudaimonistically (wellness-related) negative side-effects. My results show that it is people's sense of what agents deserve, and not the moral status of side-effects, that drives intuition.

In line with his empirically grounded theory that interpretations of other people's minds do not follow scientific principles, Knobe hypothesizes that our judgments of the intentional nature of side-effects depend on the side-effect's assumed moral status.

I conducted a study involving 150 participants which challenges this hypothesis (Brogaard 2010b). The participants were divided into four groups of 25, and the subjects in each group were randomly assigned a vignette featuring either an economically or

eudaimonistically negative side-effect or a positive side-effect[1]; plus I had two of these groups of 25 test the last vignette of the target article (hence totaling 150 subjects). Each of the first four groups received one of the following different vignettes:

(1A) The famous stand-up comedian Rob's personal assistant went to Rob and said, "We are thinking of changing your medication. It will help your popularity immensely by completely treating your stage fright, but it will also hurt you by causing morning headaches." Rob answered, "I don't care at all about having morning headaches. I just want to be as popular as possible. Let's switch to the new medication." Sure enough, Rob suffered from morning headaches.
(1B) [. . .] "It will help your popularity immensely by completely treating your stage fright, and it will also help you by curing your morning headaches." Rob was cured of his morning headaches.
(2A) The famous stand-up comedian Rob's personal assistant went to Rob and said, "We are thinking of hiring a new PR assistant. It will help your popularity immensely, but it will also harm your financial situation." Rob answered, "I don't care at all about my financial situation. I just want to be as popular as possible. Let's hire the new PR assistant." Sure enough, Rob's financial situation was harmed.
(2B) [. . .] "It will help your popularity immensely, and it will also help your financial situation." Rob's financial situation was helped.

Of the participants in the group that received vignette (1A), 84% judged that Rob intentionally harmed himself; 76% in the group receiving (1B) judged that Rob didn't intentionally help himself; 88% of the group that got (2A) judged that Rob intentionally hurt his financial situation; and 76% of the group that got (2B) judged that Rob didn't intentionally help his financial situation.

In these vignettes, the side-effects have no direct bearing on morality. But the vignettes are akin to Knobe's in describing an agent with undesirable personality traits. The agent is either greedy and self-centered (the chairman), or superficial (Rob).

I found that these personality traits figured in participants' answers to follow-up questions. When asked to "describe Rob's personality traits," 88% replied with one of the following words: "shallow," "superficial," "stupid," "flaky," "irresponsible," or "careless."

Participants given the vignettes (1A) and (1B) were also asked whether Rob deserved to suffer from headaches or economically, given the decision he made. Here, 98% checked the options "yes" or "leaning towards 'yes'." When asked to justify their answers ("Rob deserves/does not deserve to suffer from headaches/economically because:"), 72% of the participants who replied "yes/leaning towards 'yes' " used descriptive terms such as "superficial," "stupid," and "irresponsible."

The results indicate that the driving force behind rendering the negative side-effects in (1A) and (2A) as *intentional* is a feeling that Rob deserves to suffer because of his undesirable personality traits.

I hypothesize that whether a (moral or non-moral) negative outcome is considered intentional depends on whether the agent is believed to deserve the outcome or (moral or non-moral) blame associated with it. If the agent is greedy, selfish, or superficial, he is thought to deserve the bad outcome or the blame. Consequently, the outcome is considered intentional.

This hypothesis explains why the chairman in Knobe's original cases is judged to have intentionally harmed the environment but not to have intentionally helped it. Because the chairman is considered greedy and selfish, he is thought to deserve potential blame associated with harming the environment. Accordingly, the outcome is considered intentional.

To further test this hypothesis, I arbitrarily assigned one of two other vignettes, similar to Knobe's in the target article, to 50 participants:

(3A) 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 the environment, and it will also hurt our profits." The chairman of the board answered, "I don't care at all about profits. I just want us to help the environment. Let's start the new program." They started the new program. Sure enough, the company's profits decreased.
(3B) [. . .] "It will help the environment, and it will help us increase profits." . . . Sure enough, the company's profits increased.

Here, 92% said the chairman in (3A) did not intentionally hurt the company, and 72% said the chairman in (3B) did not intentionally help the company.

The first result confirms our hypothesis. The chairman in (3A) has desirable personality traits: He cares about the environment, not profits. So, the subjects infer that he does not deserve the potential blame associated with having intentionally brought about a decrease in profits, and hence, that he did not intentionally bring about the side-effect.

The results in (3B) indicate that for an agent to intentionally bring about a positive side-effect, he or she must not only deserve the outcome or the potential praise associated with it, he or she must also aim at bringing it about.

In conclusion: The results of my study are in agreement with Knobe's suggestion that people's judgments of side-effects do not rely on scientific methods, but the results disagree with Knobe concerning the underlying principles driving these judgments. Knobe (2006) proposes a model for how moral assessments affect judgments of intentional action. In his original cases, we are confronted with the side-effect, *harmed environment*. We determine that the side-effect is morally bad and that the chairman showed foresight. We then employ the principle "If the side-effect is morally bad, and the agent showed either trying or foresight, then the side-effect is intentional" and infer that the chairman intentionally harmed the environment and is to blame for his behavior (see my Fig. 1).

My study suggests a different model for the attribution of intentionality. When we are confronted with a side-effect (e.g., *harmed environment*, *harmed self*, or *harmed financial situation*), we determine whether the side-effect is negative. We then assess the agent's personality in order to determine whether he or she deserves the bad outcome or the potential blame associated with it. Finally, we employ the principle "If the side-effect is negative, and the agent showed trying or foresight, and he or she deserves the side-effect or the potential blame associated with it, then the side-effect is intentional" and infer that the agent intentionally harmed the environment, him/herself, or his/her financial situation and therefore is to blame for his or her behavior (see my Fig. 2).

In a second IRB-approved[2] project involving 1,500 participants, currently in progress (Brogaard 2010a), we seek to determine the correlation among positive side-effects, undesirable personality traits/good fortune, and intentionality. Initial results indicate that an agent's bad personality traits, a history of undeserved success, or good fortune leads us to judge that the agent did not intentionally bring about the positive side-effect and hence does not deserve praise. In a pilot study preceding this larger project, participants were presented with vignettes featuring positive side-effects but differing in terms of whether the agent had good or bad personality traits or had a history of undeserved success or failure. Agents with bad personality traits or a history of undeserved success were judged not to have intentionally brought about the positive side-effect, whereas the opposite was true for agents with good personality traits or a history of undeserved failure.

Figure 1 (Brogaard). Knobe's model of the mechanisms of the side-effect asymmetry. The identification of a morally bad side-effect triggers a selective search for features that are sufficient to judge the side-effect as being brought about intentionally.

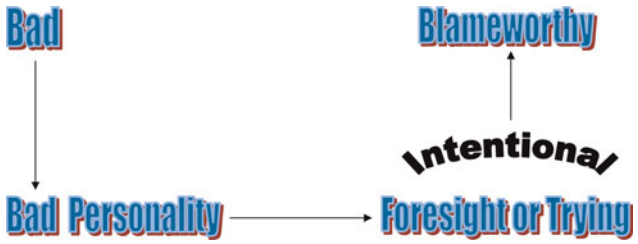

Figure 2 (Brogaard). New model of the mechanisms of the side-effect asymmetry. The identification of a morally bad side-effect triggers an assessment of personality traits, and the identification of undesirable personality traits triggers a search for features that are sufficient to judge the side-effect as being brought about intentionally.

NOTES

1. Overlapping material has been omitted.

2. This project was approved by the University of Missouri–St. Louis Institutional Review Board for the Protection of Human Subjects in Research on May 21, 2010.

The social origin and moral nature of human thinking

doi:10.1017/S0140525X10001706

Jeremy I. M. Carpendale,[a] Stuart I. Hammond,[a] and Charlie Lewis[b]

[a]*Department of Psychology, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada;* [b]*Department of Psychology, Fylde College, Lancaster University, Bailrigg, Lancaster LA1 4YF, United Kingdom.*

jcarpend@sfu.ca **shammond@sfu.ca**
c.lewis@lancaster.ac.uk
http://www.psyc.sfu.ca/people/faculty.php?topic=finf&id=67
http://www.psych.lancs.ac.uk/people/CharlieLewis.html

Abstract: Knobe's laudable conclusion that we make sense of our social world based on moral considerations requires a development account of human thought and a theoretical framework. We outline a view that such a moral framework must be rooted in social interaction.

According to Knobe, moral considerations are an integral part of the way we make sense of and reason about our social world. The problem is that Knobe requires an account of the nature of human thought explaining its moral nature, which, we argue, must be developmental (Carpendale & Lewis 2004). We take over where Knobe stops short of such a proposal, and sketch in an account of the development of thinking, showing how this is rooted in social interaction, which is moral in nature.

We propose a socially based view of the evolution and development of thinking. From this perspective, human cognition involves "moral considerations" because it originates as a social process that is gradually mastered by individuals. This social process has moral preconditions. We make sense of our social world in moral terms because this is a fundamental aspect of our human form of life, involving coordinating our actions and interests with others.

Knobe's comparison of the person as scientist versus moralist constructs a straw man (Gellatly 1997), which does not explain how moral thinking is possible, let alone how "scientific" thinking and moral thinking fit together. For us, the problem is how the person as moralist could come into being. We draw on Mead's (1934) account of the social origin of thinking and mind, according to which meaning arises interpersonally as persons come to realize the significance of their actions for others. Thus, meaning is necessarily social because it requires experiencing others' attitudes to one's actions.

The view that reflective thought originates as a social process has many implications for the role of morality in thought. At the most basic level, social interaction involves moral preconditions of responsiveness, of give-and-take, and of turn-taking (Turnbull 2003). Although we often think of morality at a grand scale, in terms of life-and-death issues, morality is also embedded in various aspects of everyday interaction with others and the way we treat one another. There is a level of morality even at the level of interpersonal interaction; it is built into the foundations of what makes our interaction possible. This social process is based on responding to one another and it is therefore moral in its roots, because we treat each other as persons, not things. We respond to one another; not to do so is to be morally accountable. At another level, human forms of communication function through assuming cooperation because we infer meaning from what others say based on the assumption that they are cooperating with us and they want us to understand them (Grice 1975a).

Another aspect of the moral preconditions of social interaction is Winch's (1972) point that, "the social conditions of language and rationality must also carry with them certain fundamental *moral* conceptions," and "a norm of truth-telling is a moral condition of language" (Winch 1972, pp. 60–63, emphasis in original). Holiday (1988) also argued that the fabric of human communication is built on the assumption that we tell the truth. Of course, lying is possible, but it is only possible because truth-telling is the default expected pattern. We develop communication in parent-child interaction in relationships of trust. It is within such relationships that communication arises and this is mastered as a form of thought.

Piaget's seminal work focused on the origins of morality in children's practical interaction with each other, and how young children instantiate reciprocity in their play (Piaget 1932/1965). Children then gradually become aware of this level of morality on which their interaction is based, and this becomes available for reflective thought; but, for Piaget, this moral understanding is rooted in earlier, practical understanding developed with social interaction.

We have outlined how the social process, which is the cradle for human forms of reflective thought, has moral foundations. Moral considerations are part of the way we make sense of our social world because thinking is rooted in, and built on, the social process, which has moral preconditions. Knobe disregards a whole tradition according to which thinking is rooted in a system of socially embedded processes of which morality is an integral part. Drawing on this tradition would enable Knobe to dispense with the view of the person as a scientist and instead consider thinking as embedded in and emerging from social interaction, which has moral preconditions at a number of levels.

Moral evaluation shapes linguistic reports of others' psychological states, not theory-of-mind judgments

doi:10.1017/S0140525X10001718

Florian Cova,[a] Emmanuel Dupoux,[b] and Pierre Jacob[a]

[a]*Institut Jean Nicod, Ecole Normale Supérieure, 75005 Paris, France;* [b]*Laboratoire de Sciences Cognitives et Psycholinguistique, Ecole Normale Supérieure, 75005 Paris, France.*

florian.cova@gmail.com
emmanuel.dupoux@gmail.com
jacob@ehess.fr
piotrjacob@gmail.com

Abstract: We use psychological concepts (e.g., *intention* and *desire*) when we ascribe psychological states to others for purposes of describing, explaining, and predicting their actions. Does the evidence reported by Knobe show, as he thinks, that moral evaluation shapes our mastery of psychological *concepts*? We argue that the evidence so far shows instead that moral evaluation shapes the way we *report*, not the way we think about, others' psychological states.

Knobe has shown that people are far more likely to judge that an agent intentionally caused (or had the desire to cause) a negative side-effect than a positive one (e.g., harming vs. helping the environment). In his target article, he argues that such asymmetries are good evidence for a "moralist" (as opposed to a "scientific") picture, according to which the naïve human capacity to ascribe psychological states to others for the purpose of describing, explaining, and predicting their actions presupposes the moral cognitive capacity to evaluate and judge others. He also offers an interesting semantics of psychological predicates such as *intention, deciding, desiring.* We think that neither Knobe's evidence nor his semantic analysis supports the moralist picture.

Knobe's semantics for psychological predicates can be seen as an extension of the semantics of gradable predicates such as *cold*. Following Pettit and Knobe (2009), suppose a beer and a coffee are both at the temperature of 20°C. Application of "cold" might plausibly yield a true statement in the coffee case and a false statement in the beer case. People rate each liquid relative to a default value that specifies what it is supposed to be like for it to be cold. In other words, the concepts respectively expressed by the words "coffee" and "beer" generate different standards of comparison for the application of the predicate "cold." Similarly, the concepts expressed, respectively, by "harm" and "help" generate different comparison classes for the application of psychological predicates (e.g., "desire," "intention"). The threshold generated by the concept *harm* is significantly lower than the threshold generated by the concept *help*. As a result, people are more likely to judge, for example, that the chairman had the desire to harm than the desire to help the environment.

Now, the fact that the semantics of gradable predicates can be extended to psychological predicates is *not* convincing evidence for the moralist picture of naïve psychology. Does the fact that the concepts *harm* and *help* generate different *moral* standards for the application of psychological *predicates* show that our understanding of psychological *states* itself is driven by moral evaluations? Consider the standards involved in the application of the quantifier *many*. Suppose that five children died in a fire and five children survived. The concepts expressed by "die" and "survive" generate different standards for the application of one and the same quantifier "many." When asked, most people were inclined to accept that many children died, but to deny that many survived. But it would be odd to conclude, on this basis, that our mastery of numerical concepts expressed, for example, by the quantifier "many" (our numerical cognition) is shaped by moral evaluation.

In fact, Knobe's own semantics for psychological predicates is not consistent with the assumption that normative standards and moral evaluation directly shape our mastery of the relevant *concepts* of psychological states. On his account, a speaker's assumption about an agent's "*pro*-attitude" towards either a negative or a positive outcome will change the speaker's willingness to apply a psychological predicate (e.g., "intentional") to the agent's action. But if so, then the psychological concept of an agent's pro-attitude must be retrieved and used by the speaker before moral considerations can come into play.

We have started to address the empirical question whether moral evaluation shapes, not just our application of psychological predicates, but our very understanding of mental states themselves. Instead of testing the distinct conditions in which participants are willing to apply the verb "desire" for evaluating the chairman's action, we asked them to use their psychological concept in order to *predict* the chairman's decision. We designed such an experiment and ran the following study on 40 participants. After receiving either the HARM or HELP scenario, participants were then given the following text:

> Imagine that, before the program is started, the VP comes back to the chairman and tells him: "It appears that we have to choose between three programs. All three will generate the same amount of benefits: hundreds of millions of dollars. The difference is that program A will have no impact on the environment, while program B will harm the environment and program C will help the environment. Anyway, it will be impossible to prove our responsibility in anything that would happen to the environment. So, if we harm the environment, no one will know of our responsibility. But, if we help the environment, that won't benefit our image. Starting program B or C will cost 10 dollars more than starting program A." If the chairman had to make this choice, what program would he choose? A, B, or C?

In this case, if participants think that the chairman has the desire to harm the environment, then they should select answer B. If they think that he has the desire to help the environment, then they should select C. If they take the chairman to be indifferent to the environment, then they should select A. Now, if we assume that their inclination to apply the verb "desire" is a reliable guide to their prediction of the chairman's choice, then we should make the following prediction: Reading the HARM scenario should cause participants to select B more than reading the HELP case should cause them to select C.

Among the participants who received the HARM case, 90% answered A, 0% answered B, and 10% answered C. Among those who received the HELP case, 80% answered A, 10% answered B, and 10% answered C. Clearly, the participants' predictions show that they do not think that the chairman's desire to harm the environment, in the HARM case, is stronger than the chairman's desire to help the environment, in the HELP case. Arguably, the moral standards triggered, respectively, by the concepts expressed by "harm" and "help" generate different comparison classes for the application of the verb "desire," which might enable people to convey to others their moral opinion of the chairman. But even so, these moral standards did not affect participants' use of the psychological concept *desire* in the prediction of the chairman's choice.

Qualitative judgments, quantitative judgments, and norm-sensitivity

doi:10.1017/S0140525X1000172X

Paul Egré

Ecole Normale Supérieure, Département d'Etudes Cognitives, Institut Jean-Nicod (ENS-EHESS-CNRS), 75005 Paris, France.

paul.egre@ens.fr **http://paulegre.free.fr**

Abstract: Moral considerations and our normative expectations influence not only our judgments about intentional action or causation but also our judgments about exact probabilities and quantities. Whereas those cases support the competence theory proposed by Knobe in his paper, they remain compatible with a modular conception of the interaction between moral and nonmoral cognitive faculties in each of those domains.

Joshua Knobe makes three main claims in his paper. The first is that the influence of moral considerations on our judgments does not appear to be limited to the concept *intentionally*, nor even to closely related concepts such as *intention* and *intending* (sect. 3.2). Thus, it appears to affect our judgments about causation, knowledge, desire, and a number of other attitudes or processes. Knobe's second main claim is that the asymmetry found by Knobe and colleagues in people's judgments for such cases depends essentially on our normative evaluation with regard to counterfactual actions or situations; namely, on what should or could have been the case. Knobe's third claim and fairly radical conclusion, finally, is that we cannot make "a clear division

whereby certain psychological processes are devoted to moral questions and others are devoted to purely scientific questions" (sect. 6, last para.).

In this commentary I would like to add further evidence in support of Knobe's first two claims, but express why I think we should be skeptical of the main conclusion he draws from them.

In agreement with Knobe's first claim, it may be pointed out that moral considerations influence at least two other general competences that would appear prima facie to be non-moral and that are not mentioned in Knobe's paper; namely, our qualitative evaluation of precise numerical *probabilities* and our qualitative evaluation of precise *quantities* (Egré 2010). The evaluation of identical numerical probabilities is known to be subjectively influenced by how detrimental the outcome is perceived. The effect has been called the *severity bias* in the psychological literature (see Bonnefon & Villejoubert 2006; Pighin et al. 2009; Weber & Hilton 1990).

For example, Pighin et al. (2009) ran an experiment comparing the evaluations made by four groups of pregnant women of a scenario in which a gynecologist tells Elisa, a 30-year-old pregnant woman, that "there is a risk of [1 in 28; 1 in 307] that your child will be affected by [Down's syndrome; insomnia]." Subjects in each group were asked to rank the probability communicated for each disease on a 7-point scale ranging from "extremely low" to "extremely high."

What Pighin et al.'s study found was that when the numerical risk for the two conditions was made the same, the women still ranked the probability of the child getting Down's syndrome as significantly higher than for insomnia. Even the probability of 1/307 for the child getting Down's syndrome was ranked higher than the probability of 1/28 for insomnia. Moreover, subjects were asked to rank each disease according to how severe they judged it to be. Their assessments of probabilities were found to correlate with those severity judgments.

Cova and Egré (2010) looked for the same effect regarding people's qualitative evaluation of identical quantities in terms of the word *many*. Subjects were given a scenario reporting that a fire had broken out in a school in which there were 10 children, 5 of whom died in the fire and 5 managed to escape. Each subject had to judge true or false the two sentences: "Many children perished in the fire. Many children survived from the fire." Irrespective of the order in which the sentences were presented, the vast majority of subjects agreed that many children had perished; but they did not agree that many children had survived, despite the identical quantities and ratios involved.

Such cases comport with Knobe's model and main explanatory hypothesis in his paper (see also Pettit & Knobe 2009); namely, they suggest that our subjective evaluation of probabilities or quantities, just like our evaluation of causation or intentional action, is sensitive not only to extensive magnitudes or processes, but also to *normative expectations* that are highly context-dependent and that vary with the kind of outcome under consideration.

For example, it is known from the semantics literature that our judgments concerning whether *many As are Bs* are not purely *extensional* (see Fara 2000; Lappin 2000; Sapir 1944). That is, as the data with Cova confirm, those judgments do not merely depend on the cardinality of As and Bs and on the ratio of As to Bs; they *intensionally* depend on the kind of entities referred to by A and B, and on what is taken to be either normal or more desirable relative to context.

In agreement with Knobe's remark about the importance of counterfactual evaluations, presumably we judge that *many children died* because we reason that in a better and alternative course of events, *fewer children would have died* (and as a result, that *more would have survived*). Similarly, how high a probability value is considered for an outcome may depend on how much more probable or less probable we consider that outcome could be or should have been.

It would be quite doubtful, however, to infer from those considerations that we cannot distinguish between the moral processes that influence our qualitative evaluation of quantities or probabilities and the non-moral processes that underlie our scientific judgments based on numerical quantities or probabilities. Indeed, when it comes to having a scientific attitude towards relative or absolute quantities, our evaluation can safely rely on our nonmoral capacity to compare extensional magnitudes. (Contrast "Did *many* children die?," which calls on our subjective and moral evaluation, with "*How many* children died?," which can be given an exact and objective answer.)

More generally, I wish to make the qualification that whereas sensitivity to normative expectations is most likely directly encoded in the *lexical semantics* of most of our *qualitative* vocabulary (see Egré 2010; Kennedy 2007), including for vague concepts such as *knowing*, *desiring*, *causing*, and so on, this remains compatible with the hypothesis Knobe appears to reject in his article. That is, it is compatible with the view that our *cognitive competence* in each of those domains works in a modular way, based on the interaction of non-moral evaluative faculties and moral evaluative faculties.

From Knobe's interesting data and examples, it would be safer to conclude that our folk concepts of causation, knowledge, and desire are irreducibly norm-sensitive, without that impugning the division between moral and non-moral cognition.

ACKNOWLEDGMENT

Research supported by the Agence Nationale de la Recherche (grant ANR-07-JCJC-0070). Thanks to F. Cova, S. Pighin, D. Ripley, and P. Schlenker for exchanges related to this commentary.

Modalities of word usage in intentionality and causality

doi:10.1017/S0140525X10001731

Herbert Gintis

Santa Fe Institute and Central European University, Northampton, MA 01060.

hgintis@comcast.net **http://people.umass.edu/gintis**

Abstract: Moral judgments often affect scientific judgments in real-world contexts, but Knobe's examples in the target article do not capture this phenomenon.

Moral considerations often affect reasoning about facts in the real world, clouding the judgments of both scientists and non-scientists. The elementary psychological processes that underlie this phenomenon are important to uncover. The experimental evidence presented in Knobe's target article, however, does not illuminate these underlying judgments.

Consider first the scenario in which a profit-maximizing individual A chooses an action that harms versus helps the environment, and a majority of subjects say the harm was intentional but the help was unintentional. Is there a disagreement concerning the facts among decision-makers? Almost certainly not. For instance, all subjects might agree with the assertion that A foresaw the effect of his decision on the environment and did not factor in this effect in deciding upon his action. All subjects must agree with this, in fact, because the description of the situation says precisely this. It follows that attributing intentionality in one case and not the other is not a judgment of fact, but rather a moral judgment. The experiment then shows that moral judgments affect other moral judgments, which is not a contested assertion.

One might object that attribution of intentionality is a factual statement concerning an individual's mental state, and sometimes indeed this is the case. For instance, we might conclude that after copulation, an insect may "intentionally" feed himself to his mate, or that the prey may "intentionally" reveal his

awareness of the predator to the predator. In such cases, we are saying that it is a normal part of the behavioral repertoire of the organism to engage in this act even when the organism has the capacity and the information to behave otherwise. But *intentionality* has a distinct second meaning that lies clearly in the moral realm. We say a undesirable result of an individual decision is "intentional" if the individual foresaw the result and could have prevented the result and achieved all other effects of the decision, except that doing so would have incurred additional personal cost.

Intentionality has yet a distinct third meaning, also in the moral realm. We say a welcome result of an individual decision is "intentional" if the individual foresaw the result and acted to bring about the result at personal cost.

When a subject says that the harm was "intentional," it is most plausibly the second meaning that is being invoked. When a subject says that the help was "unintentional," it is most plausibly the third meaning that is being invoked. We rule out the first meaning of "intentional" in these cases – because this meaning is strictly factual, whereas the context of the situation calls for a moral evaluation.

The interpretation of this evidence is complicated by the fact that there are several other commonly used meanings of *intentionally*, one being "foresaw the result and acted in order to achieve this result." In this sense, profit-maximizer A did not intentionally harm in the first scenario and did not intentionally help in the second. Very likely, many subjects chose to use this definition, despite the fact that it renders the choice completely trivial, as the statement of the problem includes non-intentionality overtly in the description of the situation.

The Gricean analysis of meaningful communication is relevant here. According to Grice (1975b), in normal conversation, a listener assumes that when a speaker solicits information, the speaker expects the information to be useful to the speaker. Thus, if someone asks, "Is there a washroom on this floor?" acceptable answers include "Yes, down the hall on the right," or "Yes, but it is out of order; there is a working washroom the next flight up," or "You'll have to go across the street." A simple yes or no would be considered a somewhat bizarre answer. In the current case, some of the common usages of the word "intentionally" are explicitly assumed in the statement of the situation, so a Gricean subject can supply useful information only by referring to those usages of the terms that require some sort of substantive inference. These usages are the second and third ones defined above.

Related problems of the multiple meaning of words beset Knobe's causation analysis. Consider the scenario of the philosophy department receptionist and the taking of pens. The question as to whether the professor, the administrator, or both caused the problem is not a matter of fact. The facts are laid out quite clearly in the statement of the scenario, and would be agreed upon by all. The notion of "cause" in question is not that of Newtonian mechanics, but rather systems theory or product design. To see this, let us change the scenario a bit, to a machine that needs a certain level of motor oil to prevent seizure:

> Half the oil is devoted to a mechanism that burns 10% of its oil allotment each day, the lost oil being replenished at the start of each day. The other half of the oil is devoted to a mechanism designed to burn no oil at all. At the end of one day, the machine seizes up and it is determined that the first mechanism consumed its allotted 10% of oil, but the second mechanism consumed an additional 10% through a malfunction.
>
> If asked whether the first mechanism, the second mechanism, or both "caused" the failure, the correct answer is the second.

There is here, of course, no factual dispute and the inspectors are making no moral judgments in placing blame on the second mechanism. In general, when a complex mechanism fails, blame is placed on elements that failed their designed tasks, even if in some sense their behavior according to Newton's laws was exactly the same as other elements that performed as designed.

ACKNOWLEDGMENT

I would like to thank the European Science Foundation for financial support.

Morals, beliefs, and counterfactuals

doi:10.1017/S0140525X10001743

Vittorio Girotto,[a] Luca Surian,[b] and Michael Siegal[c]

[a]*Department of Arts and Design, University IUAV of Venice, 30123 Venice, Italy, and Laboratory of Cognitive Psychology, CNRS and University of Provence, 13003 Marseilles, France;* [b]*Department of Cognitive Sciences and Education, Center for Mind/Brain Sciences. University of Trento, 38068 Rovereto (TN), Italy;* [c]*Department of Psychology, University of Sheffield, Western Bank, Sheffield S10 2TP, United Kingdom.*

girotto.vittorio@gmail.com **luca.surian@unitn.it**
m.siegal@sheffield.ac.uk
http://www.iuav.it/Ricerca1/Dipartimen/dADI/Docenti/girotto-vi/index.htm
http://portale.unitn.it/cimec/persone/luca.surian
http://alacode.psico.units.it/index.html

Abstract: We have found that moral considerations interact with belief ascription in determining intentionality judgment. We attribute this finding to a differential availability of plausible counterfactual alternatives that undo the negative side-effect of an action. We conclude that Knobe's thesis does not account for processes by which counterfactuals are generated and how these processes affect moral evaluations.

Ever since Aristotle's *Nicomachean Ethics*, there has been debate over the extent to which there is separation between morality and cognition. We applaud Knobe's modern effort to integrate the investigation of these areas. There are three main reasons, however, to doubt his thesis according to which moral evaluations affect the ordinary understanding of social and psychological phenomena "from the very beginning" (target article, sect. 5.3, para. 3).

First, moral and non-moral evaluations of the social world do not always work together. In particular, in support of the position that Theory of Mind reasoning is not theory-like and does not proceed in terms of a process that can be characterized in terms of "child-as-scientist" (Leslie et al. 2004), preverbal infants appear to possess basic mind-reading skills (e.g., Surian et al. 2007). No evidence suggests that such acquisition depends on input from moral competencies. Moreover, children with selective impairments of mind-reading skills appear to have an intact ability to make some basic moral judgments (Blair 1996; Leslie et al. 2006b).

Second, even when moral evaluations appear to shape ordinary intuitions about the social world, non-moral considerations are a necessary input to the shaping of these intuitions. We have found that both adults (Pellizzoni et al. 2010) and preschoolers (Pellizzoni et al. 2009) attribute intentionality to a negative side-effect produced by an agent who was not aware of it. By contrast, participants did not do this when the agent was described as having a false belief about the negative side-effect. When the side-effect was positive, participants judged that it had been produced unintentionally, regardless of whether the agent believed that it could occur or not. Thus, evaluative considerations interact with belief ascription in determining intentionality judgment.

Third, counterfactual thinking affects moral evaluations, rather than vice versa. We have attributed the above-described results to a differential availability of plausible counterfactual alternatives that undo the negative side-effect. When individuals read about an agent who did not know that his action could produce a negative side-effect, they could easily think, "Had he made an inquiry, he would have discovered the side-effect and made a different choice." Indeed, when readers undo the negative outcome of a story, they alter the protagonist's choices (Girotto

et al. 1991). But when a misinformed agent had no reason to anticipate a negative side-effect, individuals could not easily imagine a plausible alternative (e.g., "Had he imagined that he was misinformed, he would have made a different choice").

Counterfactual thinking seems to play an important role in Knobe's thesis, too: Individuals attribute intentionality to a negative but not to a positive side-effect because they tend to construct alternatives that are morally right ("If the agent had chosen differently, he might have produced a positive side-effect") rather than morally wrong ("If the agent had chosen differently, he might have produced a negative side-effect"). The problem with Knobe's interpretation that moral evaluations determine the selection of counterfactuals is that it does not explain how counterfactuals are generated or how counterfactuals affect other mental activities, including moral judgment.

His interpretation neglects the finding that individuals do construct morally dubious alternatives. For example, they imagine breaking the constitutive rules of a game in order to undo a failure (e.g., Girotto et al. 2007). When applied to intentionality attribution, Knobe's interpretation appears to confuse morality with normality. According to Knobe, individuals who read the positive side-effect story do not imagine the chairman damaging the environment because this possibility is morally wrong. We would say that they don't do so simply because this possibility alters normal events, that is, the normal tendency of chairmen to seek to make profits (Kahneman & Miller 1986; Uttich & Lombrozo 2010). With regard to the effects of counterfactual thinking, Knobe's interpretation neglects the finding that moral evaluations often depend on the availability of counterfactual alternatives. For example, individuals attribute more compensation to the victim of an accident (and more responsibility to the perpetrator) when it is preceded by exceptional rather than by normal circumstances (Macrae 1992). Finally, Knobe's interpretation cannot easily explain our results: The possible alternatives evoked by the non-informed agent version (e.g., "Had he made an inquiry. . .") were not morally different from those evoked by the misinformed agent version (e.g., "Had he imagined that he was misinformed. . ."). Yet, only in the first case did individuals attribute intentionality to the negative side-effect (Pellizzoni et al. 2010).

To investigate the relations between the moral and non-moral facets of naïve psychology remains a high priority for future research. However, in this connection, it is not necessary to postulate that moral evaluations play a pervasive role in the ordinary understanding of intentional actions.

Questioning the influence of moral judgment

doi:10.1017/S0140525X10001755

Steve Guglielmo

Department of Psychology, Brown University, Providence, RI 02912.
steve_guglielmo@brown.edu
http://research.clps.brown.edu/mbq/guglielmo/

Abstract: Moral judgment – even the type discussed by Knobe – necessarily relies on substantial information about an agent's mental states, especially regarding beliefs and attitudes. Moreover, the effects described by Knobe can be attributed to norm violations in general, rather than moral concerns in particular. Consequently, Knobe's account overstates the influence of moral judgment on assessments of mental states and causality.

Knobe's "person as moralist" account provides a novel contribution to the study of human morality. Whereas most research in this domain has examined the features of behavior that guide moral judgment (Cushman 2008; Guglielmo et al. 2009; Shaver 1985) or the processes that underlie moral judgment (Greene 2008; Haidt 2001), Knobe's target article extends the literature by probing the influence of morality on other psychological judgments.

Despite its promise, however, Knobe's account has several limitations. Knobe neither measures nor defines *moral judgment*, leaving it unclear precisely what the account posits and how it may be falsified. Nonetheless, any conceptualization of moral judgment consistent with Knobe's account necessarily relies on substantial information about an agent's mental states. Moreover, the results described by Knobe are likewise obtained by instances of entirely non-moral norm violations. Finally, Knobe should clarify why it would be the case that people's moral judgments of badness and blame share no direct relationship.

The crux of Knobe's argument is that moral judgments of badness (hereafter "MJ1") impact judgments about an agent's mental states and causal role, which thereby impact moral judgments of blame (hereafter "MJ2"). Although MJ2 are often measured, studies of Knobe's account rarely (if ever) measure MJ1. It is therefore critical to know the conditions under which such judgments arise. To this end, Knobe claims to examine "judgment[s] that *the agent's action itself is bad*." But this definition does not provide much clarity – if MJ1 are not simply judgments about bad outcomes (sect. 4.1.3, para. 4), are they judgments that an agent *caused /knew about/ intended* something bad? Absent either a measurement or definition of MJ1, it is unclear precisely what is alleged to influence mental state and causality assessments, and how one could attempt to falsify the account.

In any case, Knobe's account would be most compelling if MJ1 arise in the absence of any considerations of the agent's mental states (which, after all, are proposed to be influenced by MJ1). However, this is clearly not the case. First, the agent's knowledge is relevant to these moral judgments. For example, the harming chairman's action is bad in part because he knew that harm would occur. When agents lack knowledge of the harmful consequences of their action, people no longer view the consequences as intentional (Nadelhoffer 2006b; Pellizzoni et al. 2010). According to Knobe's account, therefore, such actions must not be bad. But if this is true, then MJ1 require consideration of an agent's knowledge.

An agent's attitude is likewise relevant to moral judgment. The harming chairman's action is bad in part because he displayed absolutely no concern for the environment. When an agent regrets or feels bad about a negative outcome, people are markedly less likely to say the action was intentional (Cushman & Mele 2008; Guglielmo & Malle, in press; Phelan & Sarkissian 2008). On Knobe's account, therefore, such actions also must not be bad, suggesting that MJ1 require consideration of an agent's attitude. Accordingly, MJ1 substantially depend on mental state information, particularly regarding beliefs (that the agent know about a negative outcome) and attitudes (that the agent not care about the outcome). These two elements are widely recognized as essential inputs to moral judgment (Cushman 2008; Darley & Shultz 1990; Guglielmo et al. 2009; Young & Saxe 2009).

Even if one grants that morality impacts mental state judgments, this effect appears to be a special case of norm violation more generally. In fact, the same empirical patterns on which Knobe's account is based are also found for cases of norm violations that have nothing whatsoever to do with morality (Machery 2008). For example, people judged an agent's making of black toys to be more intentional when doing so violated, rather than conformed to, the conventional color designation (Uttich & Lombrozo 2010). People also judged it more intentional to violate a dress code than to conform to one (Guglielmo & Malle, in press). This is because norm violations – whether moral or not – provide diagnostic information about a person's disposition, motives, intentions, and so on (Reeder & Brewer 1979; Skowronski & Carlston 1989). Interestingly, Knobe's recent work adopts precisely this explanation, highlighting the impact of non-moral norms on causality judgments (Hitchcock & Knobe 2009). But this perspective suggests that people are not

"moralists" at all; rather, their judgments are sensitive to norms, just as those of a "scientist" would be.

Setting aside the criticisms raised here, Knobe should clarify a puzzling aspect of his proposed account. Knobe distinguishes between early MJ1 (*badness* judgments) and later MJ2 (e.g., *blame* judgments). One might expect these judgments to be tightly linked, as they both assess the morality of a given action. However, the connection between them is argued to be fully mediated by non-moral assessments (e.g., regarding mental states and causality, see Figure 5 of the target article). Knobe's account would benefit from a psychological explanation for the existence of such a circuitous path between the conceptually similar MJ1 and MJ2. Why might it be that two moral judgments have no direct relationship to each other?

One possible answer to the puzzle is that MJ1 are not actually *moral* judgments, but simply judgments about whether an action violated an expectation. Such expectations are sometimes a function of valence – people expect others to bring about positive events and avoid negative ones (Pizarro et al. 2003). Perceivers may adopt different thresholds for what constitutes a relevant mental state or causal role, depending on the extent to which the action violates expectations. This possibility is largely consistent with Knobe's discussion of default attitude positions (Figs. 8 and 9 of the target article), except that Knobe maintains the threshold is set by moral judgments in particular. Given the discussion here, it is not clear how this can be true. Although expectations (including, but not limited to, those concerning valence) may impact the evidential threshold set by perceivers, moral judgments depend on assessments of an agent's mental states. Accordingly, the claim that such assessments are "suffused with moral considerations" (sect. 5.3, para. 3) is greatly overstated.

Person as lawyer: How having a guilty mind explains attributions of intentional agency

doi:10.1017/S0140525X10001767

Frank Hindriks

Faculty of Philosophy, University of Groningen, 9712 GL Groningen, The Netherlands.

f.a.hindriks@rug.nl

http://www.rug.nl/staff/f.a.hindriks/index

Abstract: In criminal law, foresight betrays a guilty mind as much as intent does: both reveal that the agent is not properly motivated to avoid an illegal state of affairs. This commonality warrants our judgment that the state is brought about intentionally, even when unintended. In contrast to Knobe, I thus retain the idea that acting intentionally is acting with a certain frame of mind.

The experimental findings Knobe discusses suggest that normative considerations influence our judgments about non-normative issues. The core finding is this: When an individual brings about a harmful side-effect, foresees that he does so, but does not care about it, people nevertheless tend to judge that he does so intentionally. The key question is whether these judgments are correct, calling for a revision of prevailing analyses of intentional action, or whether no such revision is needed since the judgments are simply incorrect. The controversy surrounding this finding is marked by a conspicuous absence of the legal perspective (Malle & Nelson 2003 and Nadelhoffer 2006a are exceptions). In particular, it has gone unnoticed that the way *intent* and *foresight* are interpreted in law provides support for taking the attributions of intentional agency at face value.

Intention and foresight in criminal law. In criminal law, it is common practice to classify cases of foresight as intent, even though the agent did not strictly intend to bring about the relevant effect. Courts are "entitled to infer" intent from foresight (Ashworth 2006, p. 178). The underlying idea is that foresight betrays a guilty mind (*mens rea*) as much as intent does. This insight is often expressed by saying that the agent "obliquely" intends the effect. An agent obliquely intends an effect when she anticipates it as a consequence of her action, even though it does not contribute to the successful performance of that action (Duff 1996, p. 17). In practice, this notion of an oblique intention is used only in relation to harmful consequences. So an agent who is said to obliquely intend a consequence will be someone who lacks the proper motivation to avoid a harmful or illegal consequence of her action. My *hypothesis* is that this misalignment between what actually motivated her and what (she realized) should have motivated her warrants our judgment that she brought about the effect intentionally (see Hindriks 2008).

Many foreseen consequences do not really concern us. I might realize that I increase the humidity in my bathroom when I take a shower, but normally I do not really care about this and I have no reason to do so. It would be odd to say that I increase the humidity "intentionally." When a consequence of my action is harmful, however, I should be concerned about it. When a foreseen consequence should concern us, it makes much more sense to attribute intentional agency. Intentional action would then be broader than intent, but narrower than foresight. Duff argues that "the wider legal definitions of 'intention' try to capture this broader notion" (1990, p. 37). This suggests that legal practice supports the folk attributions of intentional agency.

The* mens rea *explanation. On my hypothesis, intentional action is of special interest to lawyers or prosecutors. A misalignment between what actually motivated a defendant and what (he realized) should have motivated him bears directly on whether he acted intentionally, and thereby on whether he satisfies the *mens rea* requirement of the relevant criminal offense. In light of this, I call my account of the attributions of intentional agency that Knobe has investigated "the *mens rea* explanation." This explanation shares with Knobe's account the idea that moral considerations figure in the competences of people who attribute intentional agency. The way in which Kobe's account differs from mine can be illuminated in terms of the distinction between conduct and fault, between *actus reus* and *mens rea* (interpreted broadly to cover both illegal and immoral acts). Knobe argues that the moral character or badness of the side-effect influences judgments of intentional action. This is a matter of *actus reus* rather than *mens rea*.

Knobe's *actus reus* explanation has an important drawback. The moral character of a consequence that constitutes the *actus reus* is not something mental and does not concern the motivation of the agent. Thus, the *actus reus* explanation severs or significantly weakens the tie between intentional action and motivation. In particular, Knobe has to abandon the idea that acting intentionally is a matter of acting with a certain frame of mind. This is a core commitment in our understanding of intentional action (Bratman 1987; Setiya 2003; Velleman 1989). The *mens rea* explanation places the agent's failure to be motivated appropriately at the center of the relevant attributions of intentional agency; it focuses on the agent's ignoring of a normative reason that counts against his intended action. It thereby preserves the idea that acting intentionally is a matter of acting in a certain frame of mind.

The *mens rea* explanation has at least two other virtues. Its second virtue is that it reveals why the notion of intentional action is so useful as input for judgments about criminal and moral responsibility: Culpability and blame require both (illegal or immoral) conduct and fault (*mens rea*), and the notion of intentional action serves to provide (defeasible) evidence for fault (there is no use for a notion broader than intent for beneficial consequences, because, in contrast to blame, praise requires intent; Stocker 1973, p. 60). Many have argued that

Knobe's core finding undermines this traditional conception of intentional action and responsibility. The *mens rea* explanation shows that it actually supports it.

The third virtue concerns the shifting standard that Knobe postulates with respect to which issues such as acting intentionally, deciding, and favoring are judged. By changing the focus from what is good or bad to what the agent has reason to do, the *mens rea* account makes better sense of why the default does not apply when legal or moral issues are concerned: People are held to a different standard with respect to what motivates them because (and in particular when they realize that) they have reason to behave differently. Standards shift when legal prosecution or moral criticism becomes pertinent.

Person as moral scientist

doi:10.1017/S0140525X10001779

Nicholas Humphrey

London School of Economics (Emeritus Professor). Home address: 18 Bateman Street, Cambridge CB2 1NB, United Kingdom.

humphrey@me.com **www.humphrey.org.uk**

Abstract: Scientists are generally more moral, and moralists more scientific, than Knobe suggests. His own experiments show that people, rather than making unscientific judgements about the moral intentions of others, are behaving as good Bayesians who take account of prior knowledge.

Knobe's home university must be a remarkable place if, as he suggests, scientists there "typically leave [moral] questions to one side" (sect. 6, para. 2). In the wider world, science is nothing if not a moral enterprise. At the very least, scientists make a public commitment to tell the truth, to respect the rules of argument, to make their arguments open to refutation, not to cheat, and so on. Think of a scientist who is engaged in peerreviewing a colleague's work: He or she is probably using the "ethical circuits" in his or her brain in similar ways to a judge at a criminal trial. Contrary to the picture Knobe paints, I would say science is an approach to the world that could only have been developed by humans who were already constantly aware of right and wrong.

Persons as scientists ought to be moral. But persons as moralists ought to be scientific, too. Knobe claims that his experimental studies show that when people are morally engaged, they begin to think "unscientifically." Yet it can be argued, on the evidence of his own experiments, that the opposite is true.

Let's consider the chairman study. Subjects are asked to judge what the chairman's intentions were. But, it is important to note that, since subjects have only limited access to the facts, the best they can do is to make an informed guess. What Knobe then finds is that they guess, on one hand, the chairman intended to harm the environment, but on the other, he did not intend to help it. So, either way, they guess *the chairman's intentions were reprehensible*. But isn't this exactly what we might expect if the subjects are *rational guessers* who have, as it happens, been given prior reason to believe that *the chairman is a bad man*?

Knobe himself comes close to saying as much in the last paragraph of section 5.2 when he says that "before people even begin considering what actually happened [. . .] they make a judgement about what sort of attitude an agent could be expected to hold." However, what he does not seem to realise is that this is a thoroughly scientific approach. Philosophers of science widely agree that the best procedure under conditions of uncertainty is to adopt a Bayesian algorithm and calculate the probabilities of a particular outcome based on prior knowledge (see the discussion in Pugliucci 2010).

True enough, ordinary people as scientists are not equally attentive to all kinds of prior information. And when it comes to predicting the behaviour of others, there is no question that morally relevant information takes pride of place. In particular, as Cosmides and Tooby have shown, people tend to be on the alert for any evidence that another person has deliberately *broken a social contract*. Moreover, if and when people suspect this, they begin to think *all the more rationally* (see, e.g., Cosmides et al. 2010). Now, the evolved "cheater-detection mechanism," which Cosmides and Tooby have identified, would certainly be activated by news about the chairman who does not pull his weight in protecting the environment. We might, therefore, expect subjects in the experiment to be thinking particularly clearly about intentionality, causation, and so on.

No doubt the cheater-detection module plays a key role too when scientists review each other's scientific work – which is why we all do it so well. (What's that motto at Yale, where Knobe comes from? *Lux et Veritas* – "Light and Truth.")

The cultural capital of the moralist and the scientist

doi:10.1017/S0140525X10001780

Min Ju Kang[a] and Michael Glassman[b]

[a]*Department of Child and Family Studies, Yonsei University, Seoul 120-749, Korea;* [b]*Department of Human Development and Family Science, The Ohio State University, Columbus, OH 43210.*

mjkang@yonsei.ac.kr

Glassman.13@osu.edu

Abstract: In this commentary we explore Knobe's ideas of moral judgments leading to moral intuitions in the context of the moral thought and moral action debate. We suggest that Knobe's primary moral judgment and the setting of a continuum with a default point is in essence a form of cultural capital, different from moral action, which is more akin to social capital.

The idea that there is a difference between moral thought and moral action has bedeviled the study of moral development and decision making for years (Blasi 1980). At the core of the debate is the idea that individuals make very different decisions, and oftentimes show very different sensibilities, when they are judging others who have engaged in some type of transgression versus when they themselves are actively involved in an ambiguous social problem. Are these observed differences representative of some qualitative difference between moral thought and moral action, or is the difference the result of the same basic decision-making process adjusting to two very different situations and perspectives? (For example, it is a common theme in ethnography that actually being in the situation changes your perspective of the situation; Malinowski 1922.)

In his target article, Knobe never really addresses a possible division between moral thought and moral action in any overt way, relying primarily on judgment/decision-making scenarios describing the actions of a social agent to make his case that generalized judgments precede and serve as context for moral intuitions. What Knobe adds to the equation in his complex analysis of moral competencies is the idea that our intuitions concerning the intentions of an agent (and therefore the possible moral culpability of the agent) are deeply affected by primary, dynamic judgments of the generalized situation/dilemma faced by the agent – what Knobe refers to as "moral considerations."

We feel Knobe's thesis makes sense but leaves open two critical questions related to the moral thought/moral action dilemma. The first is: Where do these initial moral judgments that serve as context for intuition and further decision-making come from? Establishing an initial, complex moral judgment as the originating point of moral intuitions and decision-making in a sense

begs the question of what is driving moral decision processes. The second question we are left with is: What, if any, role would this continuum play in the moral actions of the individual? Do we use the same type of default system when we are making socially ambiguous decisions that might directly affect us and/or those around us? (For example, are there pre-judged lines that we will not cross?)

We have made the argument (Kang & Glassman 2010) that moral thought, including the type of moral judgments Knobe describes, is actually a form of what Bourdieu (1986) refers to as *cultural capital*, while moral action is a form of *social capital* (Portes 1998). The motivation and goal (Glassman 1996) of cultural capital is to signal to those around you that you are a member in good standing of the social group. It is a short-hand for the types of social interactions that allow individuals to establish affiliation through community standards. Moral judgments are one of the easiest forms of cultural capital to use to establish group membership, whether it is gossip around a community pool or the establishment of a common enemy, villain, and/or scapegoat.

We suggest that the primary moral judgment that Knobe describes is made in the service of cultural capital, and that it is more about signaling and establishing membership in a given community than "about controllability, about recency, about statistical frequency" (sect. 5.1, para. 5). The default position of what is acceptable for the businessman in Knobe's scenario example would change dramatically based on whether you were trying to signal membership and affiliation in the Chamber of Commerce or in the Sierra Club.

The setting of a continuum and establishing of a default is a form of cultural capital, and if we are on target in our thesis (Kang & Glassman 2010), it would all but disappear when individuals are engaged in collective moral action. In moral action, individuals are less concerned with establishing a signal/symbol system for long-term group maintenance and belonging than in coming together as a group to solve a critical problem. In moral action, the focus is almost completely on the problem at hand, rather than on who should be included (and excluded) from the working group. The action is integrated with the specifics of the problem to be solved, and as the common problem dissipates, so, too, does the motivation behind the group (Putnam 2001).

We see two reasons why there is little to be gained by using the primary moral judgments of generalized situations in moral action. The first is that group membership is malleable in problem solving, and placement in the group is dependent on abilities. The second reason is that problems are dynamic and shifting, and individuals who are taking action might have to continuously abandon or change their default point based on circumstances. To take a crude example, a person with a specific default position on sharing of community resources might take a very different view if he or she is placed in charge of such resources. (For example, how would the individuals in Knobe's academic example change if they found themselves being denied access to pens when they needed them? Or if their salary were dependent on maintaining a supply of pens?)

We take the real-world example of the recent British Petroleum (BP) oil spill to illustrate our point, similar to Knobe's businessman who does not care about risks to the environment. Suppose, before the spill occurred, people were asked about the intentions of the president of BP if he said the company could engage in deep water drilling without harming the environment: There would be a wide array of responses, directly based on the primary judgments Knobe discussed, but judgments used to signal community belonging. If you asked an officer in an environmental group, he or she might have set the default point for acceptable action so that the greater part of the continuum led to intuitions of morally bad intentions (e.g., being willing to drill at all, or not actively investing in sustainable energy). If you asked a politician from the Gulf Region, he or she might have set a default position with far more of the continuum devoted to a neutral position (e.g., drilling could occur as long as there were minimal safety precautions) in order to signal kinship with the oil-dependent community.

After the spill, many members of the two groups have acted together in attempting to stop the spill and reclaim the Gulf. Intuitions about good or bad intentions and the moral judgments that led to them have become secondary or even irrelevant for many working in this group, and it is considered bad form to bring them up. Ties have been established based on the need to solve the immediate problem. Once the problem has diminished, or retreats into the background, the social group will dissipate and moral judgments as cultural capital will move to the fore again. It represents a cycle of moral thought as cultural capital and moral action as social capital.

Are mental states assessed relative to what most people "should" or "would" think? Prescriptive and descriptive components of expected attitudes

doi:10.1017/S0140525X10001792

Tamar A. Kreps and Benoît Monin

Graduate School of Business, Stanford University, Stanford, CA 94305.
Kreps_Tamar@gsb.stanford.edu
monin@stanford.edu **http://www.stanford.edu/people/monin**

Abstract: For Knobe, observers evaluate mental states by comparing agents' statements with "defaults," the attitudes they are expected to hold. In our analysis, Knobe's model relies primarily on what agents *should* think, and little on expectancies of what they *would* think. We show the importance and complexity of including descriptive and prescriptive norms if one is to take expectancies seriously.

If you claimed at a dinner party to have no opinion about child abuse, you would get funny looks. In Knobe's analysis, because you should strongly oppose abuse, neutrality is tantamount to support. Similarly, expressing neutrality about women's suffrage, which our society supports, would appear sexist. Thus, observers do not take agents' claims at face value, but instead assess them relative to what Knobe calls a "default." Observers essentially convert an agent's claim to their own metric, much like converting Celsius to Fahrenheit, based on the object of judgment (e.g., helping vs. hurting the environment) and the associated "default" attitude (see our Fig. 1, Panel A).

This "default," defined in the target article as "what sort of attitude an agent could be expected to hold toward" an object (sect. 5.2, last para.), and elsewhere (Pettit & Knobe 2009) as what any reasonable person "would" (p. 597) or "should" (p. 598) think, is thus a central part of Knobe's model. In this commentary, we aim to analyze and clarify this concept, which we believe is more complex than Knobe lets on. There is much to be gained from such analysis, especially from distinguishing the *should* and *would* aspects of default expectations.

What influences people's expectations about how others behave and think? Certainly, one factor, as Knobe points out, is *personal moral judgment*: we expect people to behave in (what we ourselves believe is) a moral fashion. However, two other social factors seem at least as important as personal moral judgment in determining defaults: *prescriptive norms* (how we think the *group* believes people should act) and *descriptive norms* (how we think group members *actually* act, regardless of how they should). Personal moral judgments do not always correspond to group prescriptive norms, and the default expectation often depends on the latter, as when an agnostic, hearing an American presidential candidate publicly espousing agnosticism, sees this as a forceful anti-religion stance given American norms, even if it accords with his own views. Similarly, a

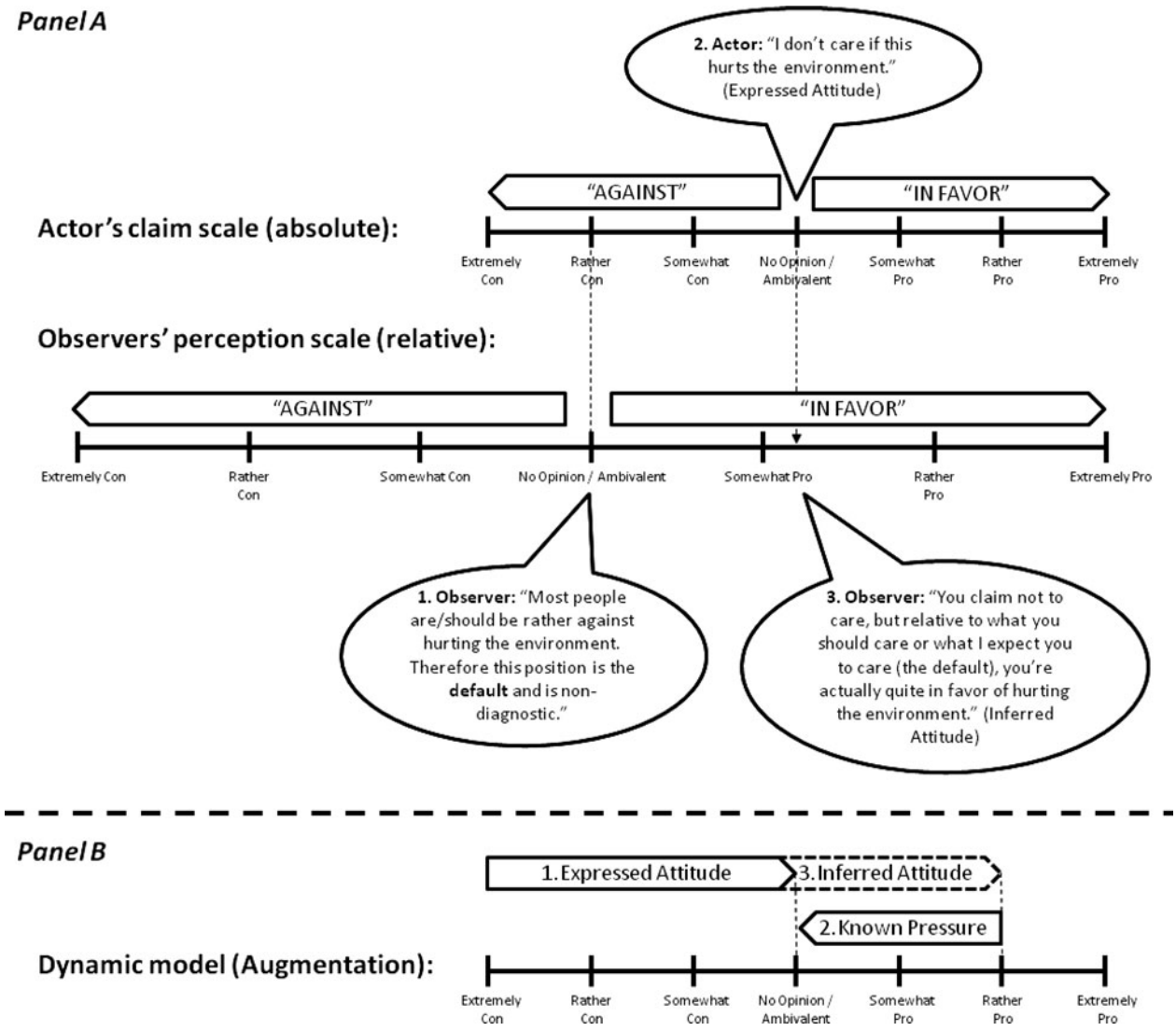

Figure 1 (Kreps & Monin). Converting expressed attitudes into inferred attitudes by reference to a default (Panel A) or to known pressure (Panel B).

default based on descriptive norms explains why, even if I know that I (personal moral judgment) and my colleagues (prescriptive norm) believe it is better to take public transportation than to drive to work, my assessment of a colleague who drives (and whether that means she "supports" public transportation) still depends on whether I know my colleagues generally drive or not.

These examples illustrate that we evaluate other people's choices not just relative to the default of our own personal values (how they *should* act), but also relative to what we can reasonably expect from others given our knowledge of the world (how they *would* act). Knobe privileges the *should* aspect: For example, one version of the pen/professor study (sect. 3.4) pits moral judgment against descriptive norms, and the relative importance of moral judgment is taken to support the model. Although Knobe's issues (the environment; reasonable rules about pens) are fairly prescriptively consensual, perceived prescriptive norms could be divorced from personal moral judgment, in which case Knobe would still favor the latter. Imagine I feel strongly that eating meat is immoral, while realizing my view is the minority one. Knobe would say that I think others who express indifference are really in favor, because my *should* default is strong opposition, even though I would not reasonably expect a random stranger to share my view (*would* default).

While Knobe may be right that *should* factors matter in many circumstances, other evidence suggests the importance of *would* factors in evaluating attitudes. For example, people use prescriptive norms to infer situational pressure and correct expectations accordingly. Observers assuming strong pressures against expressing support for harming the environment can sensibly infer a suppressed *pro*-harm attitude behind an expression of indifference (Fig. 1, panel B). Similarly, a speaker advocating immediate action by a corporation to reduce pollution is perceived as more *anti*-environment when speaking to a *pro*-environment audience, where such a message is expected, than to a *pro*-business audience (Eagly et al. 1978). Here, assumed audience pressure changes the default, although participants' personal moral judgment presumably remains constant. Ironically, this is exactly the "augmentation" process described in Kelley's (1971) attribution theory, which Knobe dismisses as a wrongheaded "person-as-scientist" theory.

Another example where *would* matters is the impact of intergroup perceptions. A devout Catholic claiming no particular opinion on Roe v. Wade might seem more in favor than a staunch feminist making the same claim. Biernat (2005) showed that expectations associated with different groups lead to such contrast effects. (Intriguingly, Biernat's research also suggests an *assimilation* effect with more objective measures – the

https://doi.org/10.1017/S0140525X10001810

Catholic would still seem less likely to get an abortion – suggesting that Knobe's might have found a different pattern using objective outcome measures.) Thus, group-specific descriptive norms evoked by agents' identities influence the default.

The value of distinguishing *should* and *would* influences on defaults is further suggested by research indicating possible interactions between them. For example, personal moral judgments affect perceived norms: Research on naïve realism and social projection (e.g., Ross & Ward 1996) shows that individuals generally believe their own judgments are rational, objective, and ethically appropriate, and therefore overestimate the similarity of others' attitudes. Also, norms can influence personal judgment: People's desire to fit in can lead them to change their own judgment to conform to perceived norms (e.g., Asch 1956). Further, descriptive norms are sometimes inferred from prescriptive norms, and vice versa (Prentice & Miller 1996).

In summary, we believe Knobe's model makes a valuable addition to our understanding of defaults and social judgments, but it seems to be unreasonably limited to factors based on "should"; for a fuller understanding of what determines people's default expectations, the model could be enriched by including other factors based on "would," such as group descriptive and prescriptive norms. Including these factors – which often have little to do with morality – might dilute the model's focus on how moral considerations suffuse social judgment, but such a change seems warranted given the important role of non-moral factors in determining default expectations. We hope future research will extend Knobe's model to include such factors.

Understanding the adult moralist requires first understanding the child scientist

doi:10.1017/10.1017/S0140525X10002037

Tamar Kushnir and Nadia Chernyak

Department of Human Development, Cornell University, Ithaca, NY 14853.

tk396@cornell.edu nc98@cornell.edu

http://www.human.cornell.edu/che/bio.cfm?netid=tk397

Abstract: Children learn *from* people and *about* people simultaneously; that is, children consider evidentiary qualities of human actions which cross traditional domain boundaries. We propose that Knobe's moral asymmetries are a natural consequence of this learning process: the way "child scientists" gather evidence for causation, intention, and morality through early social experiences.

Knobe's "person as moralist" view contests two related claims about human cognition: that it is clustered by discipline, much as university departments are, and that cognition in two "scientific" disciplines – folk psychology and causal inference – is analogous to scientific inquiry. Knobe then presents evidence that the psychology of intention and causation are "suffused with moral considerations" (sect. 5.3, para. 3), by which he means to show that there is neither a separation between disciplines, nor can reasoning about scientific topics be considered "scientific."

We suggest another perspective on these moral asymmetries: that they are, at least in part, the consequences of early links between causal learning and social learning. Specifically, they are the result of how, as children, we gather evidence for such learning by observing and interacting with people. The adult moralist recruits knowledge gained from years of social evidence gathering – years spent learning *from* people and *about* people simultaneously. Therefore, to understand the adult moralist we must first understand her predecessor – the child scientist.

For a long time, developmental psychologists studied children's knowledge separately, according to domain. Some research examined early causal reasoning – intuitions about spatio-temporal relations (Leslie & Keeble 1987; Oakes & Cohen 1990), causal mechanisms (Bullock et al. 1982; Schulz 1982), and the use of statistical cues in causal judgments (Gopnik et al. 2001; Sobel & Kirkham 2006). Other research focused on children's "mind-reading" abilities – what they knew about the intentions, desires, beliefs, and knowledge states underlying human actions (e.g., Lutz & Keil 2002; Repacholi & Gopnik 1997; Wellman 1990; Woodward 1998). Others sought to understand children's knowledge of social categories (Bigler & Liben 2007; Heyman & Gelman 2000), and still others focused on developing moral and conventional knowledge (e.g., Turiel 1983). The picture that emerged from these separate subfields is a lot like the mental university described by Knobe – separate departments for separate knowledge structures.

The domain-specific approach has lead to important discoveries about the content of early physical, biological, psychological, and social and moral knowledge. However, trying to apply this approach wholesale to learning processes has been less fruitful. Take causal learning: spatio-temporal cues and mechanism knowledge are useful, but are often unavailable. Statistical cues are also useful, but cannot help distinguish between causes and spurious correlations. Most often, ordinary causal learning depends on social interaction; evidence comes from doing things and watching others do things. Human actions are a child scientist's natural causal experiments (Gopnik et al. 2004; Schulz et al. 2007).

Importantly, along with physical evidence (e.g., toys making noise, milk spilling, sticks breaking), causal actions contain valuable social evidence (a knowing glance at the right button, a cry of "oops!", a desire for two short sticks). To evaluate the quality of causal evidence, children take knowledge, ability, and intention into account. For example, infants and preschoolers distinguish intentional actions from accidental ones, and this leads them to make different causal inferences (Carpenter et al. 1998; Meltzoff 1995). Preschoolers prefer to learn new causal relations from knowledgeable rather than ignorant causal agents (Kushnir et al. 2007). Children also treat causal evidence differently when a demonstrator is explicitly teaching them (Bonawitz et al. 2009; Rhodes et al., in press). This evidentiary link is not limited to passive observations – it influences and interacts with the evidence children generate themselves through play. Thus, when children get ambiguous evidence from another person, they privilege evidence from their own past actions (Kushnir et al. 2009), or are motivated to explore further to generate new evidence (Schulz & Bonawitz 2007).

Other research suggests that children break traditional domain boundaries when learning about people, as well. For example, infants use contingency detection (Shimizu & Johnson 2004) or violations of contiguity (Saxe et al. 2007; Spelke et al. 1995) to infer the presence of a psychological agent when other cues to agency are absent. Toddlers and preschoolers infer other people's preferences based on violations of random sampling, not merely positive regard and enthusiasm (Kushnir et al. 2010). Children may use statistical cues to track other individual regularities, such as personality traits (Siever et al, under review). They also readily track social regularities, such as norms and group characteristics (Kalish 2002; Rhodes & Gelman 2008).

From her earliest social experiences, the child scientist is engaged in a dynamic process of hypotheses formation, evidence-gathering, and theory change. The adult moralist, on the other hand, is asked to reason about a single instance of human behavior. The adult must therefore rely on her existing knowledge – knowledge acquired through this early learning process. We now have a better sense of where this knowledge begins; recent studies show early understandings of empathy, fairness, help, harm, and a host of moral precursors (e.g., Hamlin et al. 2007). Knobe's analysis encourages us not to stop with domain-specific characterizations of knowledge. Instead, we should broaden how we view evidence from human actions to include their moral and normative dimensions, and investigate how these early evidential links give rise to later moral asymmetries

in reasoning. This approach leads to interesting questions for research with adults, so long as we carefully distinguish between reasoning based on existing knowledge and the process of learning something new. When adults learn, for example, how do moral asymmetries change in response to further evidence? Is the evidence itself evaluated asymmetrically?

To conclude, while it may be wise at times to abandon the separation of disciplines, it seems premature to draw conclusions from Knobe's experimental data about the process by which they are integrated. To better understand this process, we need to look at learning at all ages, and continue research connecting moral development to both causal learning and social cognition.

Scientists and the folk have the same concepts

doi:10.1017/S0140525X10001809

Neil Levy

Florey Neuroscience Institutes, Carlton South 3053, Australia; and Oxford Centre for Neuroethics, Oxford, OX1 1PT, United Kingdom.
neil.levy@philosophy.ox.ac.uk

Abstract: If Knobe is right that ordinary judgments are normatively suffused, how do scientists free themselves from these influences? I suggest that because science is distributed and externalized, its claims can be manipulated in ways that allow normative influences to be hived off. This allows scientists to deploy concepts which are not normatively suffused. I suggest that there are good reasons to identify these normatively neutral concepts with the folk concepts.

Joshua Knobe has added considerably to our knowledge of the ways in which ordinary people attribute intentions and make judgments regarding causation. In this commentary, I do not want to criticize his claim that the competencies agents deploy in making these judgments are deeply suffused with normative influences. However, I will suggest that there are nevertheless grounds for regarding these competencies as distorting influences on our concepts. Our perfectly ordinary concept of causation (for instance), I suggest, is not normatively suffused. This is best brought out by thinking about science; I therefore begin with Knobe's claim that we ought not to understand folk judgments on the analogy of scientific hypothesis testing.

Knobe's claim that folk judgments are made in ways very unlike scientific hypothesis testing leaves us with a puzzle: Given that scientists are ordinary people too, how do *they* manage to engage in scientific research? If the relevant competencies are suffused with normative influences, how do scientists manage to free themselves of these influences (sufficiently well that they can identify them in the first place)? This question is important for several reasons, including, that if we can identify the means whereby scientists succeed in separating normative influences from the relevant judgments, we might all be in a position to make better *normative* judgments. At least on standard normative theories, our normative claims ought to follow *from*, rather than themselves cause, judgments of causation and intention; hence, separating out the normative from the non-normative might be a precondition of justified normative judgment.

So how do scientists manage to transcend the normative influences Knobe identifies? The answer is multifaceted, but an important part of it refers to the structure of the scientific enterprise. Science is an essentially distributed enterprise. The structure of a scientific community enables its members to compensate for the limitations and biases of individuals (Kitcher 1993). Individual biases can thereby be cancelled out; one scientist's bias toward a hypothesis will be cancelled out by another's against it. Of course, this cancellation process is powerless against the kind of normative influences Knobe identifies, as they are universal.

But the structure of science has a second property: it externalizes scientific knowledge. Since science, by virtue of its essentially distributed nature, requires that data and theories be available to a multiplicity of researchers, they must be presented in a format that makes this possible, and that requires externalization. Once theories and data are externalized in this way, they become available for manipulation using formal techniques, and these techniques are designed to be impervious to the normative influences Knobe identifies. They can also be manipulated through the use of methods such as double blinding, which can also serve to filter out normative influences.

One implication of the forgoing is that the finding that ordinary people are not best understood on the model of scientists is unsurprising: no one is a scientist alone. An agent can be a scientist only as part of a community of researchers engaged in systematic inquiry. The contrast between scientific judgments and folk judgments is therefore misplaced: The contrast is not between different modes of thinking so much as between different ways of manipulating mental representations; one individualistic and the other deeply social.

An important implication is that there are grounds for seeing the competencies that agents utilize in making judgments as distortions of their concepts. We do not wish to say that scientists are mistaken in making causal judgments that are not normatively suffused. We therefore should not see the concept of causation as constituted by the structure of the competencies Knobe has elegantly uncovered. Scientists are members of the folk, and their onboard competencies are identical to everyone else's, yet they understand their causal judgments, *qua* scientists, as deploying the ordinary concept of causation, not a theoretical innovation. I suspect that given the choice between the concept of causation used in science and one that is *explicitly* normative, ordinary people would also choose the former, providing further evidence that scientists use the ordinary concept.

In saying this, I take issue neither with Knobe's arguments in favor of the view that our competencies are themselves normatively suffused, nor with his correlative claim that the rival view (according to which moral judgments bias our application of our concepts) is false. I am accepting that normative influences figure into the relevant competencies, but I am claiming that nevertheless we need to distinguish between these competences and the relevant concepts, even though we probably derive the concept from the competency (via some process of idealization). The concept of causation is normatively neutral, even though ordinary people deploy the concept using competencies that are normatively suffused.

It may be that we can dissociate the normatively neutral concept from the normatively suffused competencies only by externalizing and distributing our application of our concepts. We can hope to deploy our concepts better by becoming more like scientists. Doing so does not involve changing our onboard competencies – that may be a task that is beyond us – but instead requires that we alter the context in which we deploy them. By dividing and distributing cognitive labor, and by designing institutions that filter out the normative influences, we may become better reasoners, both in the normative and the non-normativerealms.

Putting normativity in its proper place

doi:10.1017/S0140525X10001810

Tania Lombrozo and Kevin Uttich

Department of Psychology, University of California–Berkeley, Berkeley, CA 94720.
lombrozo@berkeley.edu
uttich@berkeley.edu
http://cognition.berkeley.edu/

Abstract: Knobe considers two explanations for the influence of moral considerations on "non-moral" cognitive systems: the "person as moralist" position, and the "person as [biased] scientist" position. We suggest that this dichotomy conflates questions at computational and algorithmic levels, and suggest that distinguishing the issues at these levels reveals a third, viable option, which we call the "rational scientist" position.

In this elegant and provocative article, Knobe summarizes a growing body of work suggesting that moral considerations influence a range of "non-moral" judgments, from mental state ascriptions to causal ratings. Knobe offers two interpretations for these data: (1) his preferred view of people as "moralists," and (2) the traditional position of people as intuitive "scientists," albeit poor ones subject to moral biases. We unpack these options using Marr's levels of analysis, and suggest at least one viable alternative, which we call the "rational scientist" position.

In Knobe's "person as moralist" position, "moral considerations actually figure in the *competencies* people use to make sense of human beings and their actions" (sect. 1, para. 7, emphasis added). In contrast, the "person as scientist" position claims that the "fundamental" capacities underlying these judgments are analogous to processes in scientific inquiry (sect. 2.2, para. 2). Both positions, as laid out by Knobe, involve a distinction between the "fundamental" or "primary" aspects of a cognitive system and those that are "secondary." Knobe suggests that to account for the data, the scientist approach must claim that moral considerations play a secondary role, biasing judgments that are fundamentally scientific.

Examining these positions in terms of Marr's levels of analysis (Marr 1982) reveals two different questions at play: one at the computational level, about the function of the cognitive system in question, and one at the algorithmic level, about the representations and processes that carry out that computation. For an advocate of the moralist position, the computational-level description of a cognitive system appeals to a "moralizing" function (perhaps evaluating people and their actions), and the algorithmic level is merely doing its job. For an advocate of the "biased" scientist position that Knobe considers, the computational-level description appeals to a scientific function (perhaps predicting and explaining people's actions), but the algorithmic level is buggy, with moral considerations biasing judgments.

This leaves two additional options (see Table 1). First is the "biased moralist" position, with a "moralizing" function at the computational level, but a buggy algorithm. Without a fuller computational-level analysis that provides a normative account of the judgments the algorithmic level *should* generate, this position is hard to distinguish from the "non-biased" moralist.

Second is the "rational scientist" position, which we advocate for some cognitive systems (Uttich & Lombrozo 2010). According to this position, a given cognitive system has a scientific function at the computational level, and the algorithm is just doing its job. To account for the slew of data Knobe cites, an advocate for this position must explain how moral considerations can influence judgments without threatening claims about the system's function (at the computational level) or the efficacy of the processes that carry out that function (at the algorithmic level).

In a recent paper (Uttich & Lombrozo 2010), we attempt precisely this for ascriptions of intentional action. The cognitive system in question, broadly speaking, is theory of mind: the capacity to ascribe mental states to others. Traditionally, this capacity has been conceptualized as analogous to a scientific theory, with the function of predicting, explaining, and controlling behavior. At the computational level, this puts the traditional picture in the "scientific" camp. But what are the implications for the role of moral considerations in carrying out this function? Knobe seems to assume that moral considerations have no legitimate role in this picture. But we argue the reverse: that accurately inferring mental states can in fact require sensitivity to moral considerations, particularly whether a behavior conforms to or violates moral norms.

Here, in brief, is our argument. Norms – moral or conventional – provide reasons to act in accordance with those norms. For example, a norm to tip cab drivers provides a reason to do so. Observing someone conform to this norm is relatively uninformative: We can typically infer knowledge of the norm, but not necessarily a personal *desire* to provide additional payment. In contrast, norm-violating behavior can be quite informative, particularly when other mental-state information is lacking. If we believe a person knows the norm, then observing that person fail to tip a driver suggests an underlying preference, desire, or constraint that is strong enough to outweigh the reason to conform. This same logic applies to Knobe's chairman vignettes (sect. 3.1). When the side effect of the chairman's actions helps the environment, he is conforming to a norm, and the action is relatively uninformative about his underlying mental states. When he proceeds with a plan that causes environmental harm, the action is norm violating, and allows us to infer underlying mental states that support an ascription of intentional action.

Our aim here is not to elaborate and marshal evidence for this position; we direct interested readers to Uttich and Lombrozo (2010). Rather, we hope to populate the space of possible positions and call attention to what seem to be distinct computational- and algorithmic-level assumptions lurking in the background of Knobe's target article. Knobe argues against various versions of the "biased scientist" position, but does not consider the "rational scientist" position. Like the two "moralist" positions, the biased and the rational scientist positions can be difficult to distinguish, and require a more fully specified computational-level description with a corresponding normative theory to identify which judgments stem from buggy versus non-buggy algorithms.

Knobe infuses normativity into folk considerations, painting a picture of people as moralists. But distinguishing the four positions we identify (Table 1) may actually require appeals to normativity in the generation and evaluation of empirically testable theoretical claims. In other words, we must appeal to normativity as theorists, regardless of whether or how we do so as folk. We suspect that Knobe avoids this framing as a side effect of other commitments and a preference for process-level theorizing. Whether or not it was intentional, we think it is a mistake to collapse computational and algorithmic questions. We hope future debate can restore normative questions to their proper place in scientific theorizing, whether the folk are ultimately judged scientists or moralists.

Table 1 (Lombrozo & Uttich). *Four possible positions to account for the data Knobe cites demonstrating an influence of moral considerations on non-moral judgments, such as mental state ascriptions and causal ratings. The positions are expressed in terms of Marr's levels of analysis, with one of two computational level functions, and algorithms that generate the judgments they do either as a result of their computational level functions (non-buggy) or because they are biased by other (e.g., moral) considerations (buggy).*

Four positions to account for the data Knobe cites		**Computational Level Function**	
		Scientific	**Moralizing**
Algorithm	**Buggy**	Biased Scientist	Biased Moralist
	Non-buggy	Rational Scientist	Moralist

Expectations and morality: A dilemma

doi:10.1017/S0140525X10001822

Eric Mandelbaum[a] and David Ripley[b]

[a]*Faculty of Philosophy, The Future of Humanity Institute, University of Oxford, Oxford, OX1 1PT, United Kingdom;* [b]*Institut Jean Nicod, DEC-ENS, 75005 Paris, France.*

Eric.Mandelbaum@philosophy.ox.ac.uk **davewripley@gmail.com**
http://www.fhi.ox.ac.uk/our_staff/research/eric_mandelbaum
http://sites.google.com/site/davewripley

Abstract: We propose Knobe's explanation of his cases encounters a dilemma: Either his explanation works and, counterintuitively, morality is not at the heart of these effects; or morality is at the heart of the effects and Knobe's explanation does not succeed. This dilemma is then used to temper the use of the Knobe paradigm for discovering moral norms.

Knobe presents two kinds of theories that compete with his own: motivational bias theories and conversational pragmatic theories. He presents his own theory as a competence account. While we agree with his criticisms of the other accounts, we think his taxonomy is incomplete. We would like to suggest a different form of competence account, one that does not take morality as such to play a crucial role in these effects. (In this regard, we agree with Phelan and Sarkissian [2008], Machery [2008], and indeed even Knobe and Mendlow [2004].) On our account, the effects of morality are a piece of a larger puzzle: Morality affects judgments of intentionality and related concepts only in virtue of its effects on expectations. Consequently, we think that anything affecting expectations will produce effects similar to those produced by moral norms. In fact, Knobe's own account points to a similar conclusion, although he doesn't acknowledge this.

According to Knobe's competence theory, people's moral norms influence their default expectations of others' intentions, beliefs, values, causal roles, and so on, and these default expectations in turn affect participants' judgments. Thus, he concludes, morality plays a deep role in explaining judgments in these various domains. But in this explanation, expectations are doing all the work; moral expectations have their effects only because they are *expectations*, not because they are moral. Thus, if Knobe's theory is right, we should find effects similar to the effects cited here in cases that have nothing to do with morality, but instead involve participants' non-moral expectations in parallel ways. And if this is right, it suggests that there is nothing specifically moral going on in the cases Knobe cites. These effects are, rather, effects of expectation, and expectations can be affected by both moral and non-moral factors (e.g., we expect people to have *con*-attitudes towards losing a game, although losing a game is not, normally at least, moral in any way).

Consequently, we think Knobe encounters a dilemma: Either his explanation of the effects he cites is correct, and then there is nothing especially moral at play here, but only an effect of expectations in general; or else his explanation of the effects is incorrect (in which case there may still be room for morality to play a distinctive role). Either way, Knobe finds himself in an awkward position; it doesn't seem that his explanation of morality's effects is compatible with the conclusion that moral considerations as such figure in our folk-psychological competence.

But we do not merely mean to present the dilemma. We take sides. We think Knobe's explanation is substantially correct, and that the effects Knobe finds would follow from any expectations participants hold firmly enough, whether or not those expectations have a moral character. To see whether this is indeed the case, it is not enough to look at cases that involve moral factors. Similar cases involving non-moral norms must be constructed and tested.

As a step in this direction, we have conducted some preliminary studies involving variations on the CEO cases that involve non-moral norms. These studies were conducted using participants on Amazon.com's Mechanical Turk website. We ran multiple studies attempting to measure possible non-moral effects on judgments of intention. In one such study, we used the following vignettes[1]:

Normal case:

> Two people are playing chess.
> One of them considers moving her queen to square A6. She thinks, "If I move my queen to square A6, I will capture my opponent's rook. But I don't care at all if I capture my opponent's rook; I think moving my queen to square A6 will allow me to checkmate in three moves."
> She moves her queen to square A6. Sure enough, she captures her opponent's rook.

Did she "intentionally" capture her opponent's rook?

Abnormal case:

The abnormal case was identical, except that the player allowed her own rook to be captured, instead of capturing her opponent's rook. We expected that participants reading the abnormal case would be more likely to judge that the side-effect of the player's move was intentional, when compared to participants reading the normal case. After all, it is normal to want to capture an opponent's rook, and normal to want one's own rook to remain uncaptured.

Although our results almost invariably trend in the expected direction, none actually reaches significance. (The closest result to significance arose from the vignettes given above; here, $\chi^2(1, N = 124) = 3.03$, $p = .08$.) For comparison, we also reproduced the original CEO cases using Mechanical Turk participants. Here, the results were highly significant: $\chi^2(1, N = 33) = 14.73$, $p < .001$. One possibility is that moral norms have a stronger effect on participants' expectations than do non-moral norms (or at least the non-moral norms we tested). Another possibility is that Knobe's explanation, which depends entirely on expectations, needs revision. Of course, either way, more systematic research is needed.

Our main point: One cannot only examine moral norms when judging whether Knobe's data show an effect of morality. We must look at non-moral norms as well, to find just how broad the phenomenon is. In fact, Knobe has, in the past, thought similar things. In Knobe and Mendlow (2004), the authors propose that the kind of badness that affects intentional action judgments extends beyond just moral badness. They propose this in light of studies that seem to show similar effects involving clearly non-moral factors.

These theoretical possibilities matter for further work involving this effect. If indeed the effects Knobe finds are not specific to moral norms, then we must be careful not to interpret the effects as telling us about participants' moral norms. For example, Inbar et al. (2009) use participants' judgments of intentionality as a way to measure implicit moral norms. This is risky; although judgments of intentionality might tell us something about participants' expectations in general, they cannot tell us which of those expectations are particularly moral and which are not. Use of intentionality judgments to measure implicit moral norms thus runs the risk of seeing moral norms where there are none.

NOTE

1. We thank Jesse Prinz, whose suggestion inspired these cases.

Norms, causes, and alternative possibilities

doi:10.1017/S0140525X10001834

Peter Menzies

Department of Philosophy, Faculty of Arts, Macquarie University, North Ryde, NSW 2109, Australia.

Peter.Menzies@mq.edu.au
http://www.phil.mq.edu.au/staff/menzies.htm

Abstract: I agree with Knobe's claim in his "Person as Scientist, Person as Moralist" article that moral considerations are integral to the workings of people's competence in making causal judgments. However, I disagree with the particular explanation he gives of the way in which moral considerations influence causal judgments. I critically scrutinize his explanation and outline a better one.

Knobe's general explanation of the way in which moral considerations influence intuitive judgments goes like this: In judging causation, doing/allowing, intentional action, and so on, people select alternative possibilities to compare with what actually happens and their selection of these possibilities is influenced by their moral judgments. How does this idea explain the data about people's causal judgments? Unfortunately, Knobe offers only the briefest hint in his Note 5, which suggests that moral considerations affect people's causal judgments by influencing which counterfactuals of the form "If event *c* had not occurred, event *e* would not have occurred" they regard as true. This suggested explanation doesn't work, however, for his own example in which Professor Smith's action rather than the administrative assistant's is regarded as the cause of a problem. This difference is not reflected in any difference in the counterfactuals people regard as true, since it is true that there wouldn't have been a problem if either Professor Smith or the administrative assistant hadn't taken a pen.

Luckily, Hitchcock and Knobe (2009) provide the missing elements of the explanation. Hitchcock and Knobe appeal to the finding in the literature on counterfactual availability that people are very inclined to entertain counterfactual hypotheses about what would have happened if a normal event had occurred instead of an abnormal one; and, by contrast, they are much less inclined to entertain counterfactual hypotheses in which normal events are replaced by abnormal ones. So people are willing to entertain the counterfactual about what would have happened if Professor Smith hadn't taken a pen because it "mutates" an abnormal event into a normal event. By contrast, people are less willing to entertain the corresponding counterfactual about the administrative assistant's action because it does not involve the privileged kind of "mutation." Finally, by positing that people's willingness to make a causal judgment "*c* caused *e*" goes hand-in-hand with their willingness to entertain the counterfactual "If *c* had not occurred, *e* would not have occurred," they explain why people are more inclined to regard Professor Smith as the cause of the problem.

I suspect this explanation cannot be right for two reasons. The first is that the explanation involves an uneconomical hypothesis about the capacities involved in causal cognition. The explanation implies that people have an underlying competence for understanding counterfactuals that is linked to their understanding the objective core of the causal concept (the "causal structure" in Hitchcock & Knobe 2009). This competence is exercised when people understand counterfactuals of all kinds, including the counterfactuals about Professor Smith and the administrative assistant. Sitting alongside this competence, the explanation implies, is a psychological tendency to entertain some counterfactuals as "available," a tendency aligned to people's propensity to select certain events as salient causes. This hypothesis strikes me as implausible because of its doubling up of capacities involved in causal cognition.

My second reason for suspecting that this explanation can't be correct is that empirical evidence casts doubt on the assumption that people's causal judgments depend on their counterfactual judgments. Mandel and Lehman (1996), Mandel (2003), and Byrne (2005) cite experimental data that show that people's causal judgments "*c* caused *e*" are dissociated from their counterfactual judgments "If *c* had not occurred, *e* would not have occurred": the former go with judgments about sufficient conditions and productive mechanisms, whereas the latter go with judgments about enabling conditions and preventative mechanisms.

There is another way of developing Knobe's general idea that moral considerations influence people's causal judgments by way of their selection of alternative possibilities. In their classic work, Hart and Honoré (1985) argue that the concept of actual causation originates in the situation in which a human action intervenes in the normal course of events and makes a difference in the way these develop. "The notion, that a cause is essentially something which interferes with or intervenes in the course of events which would normally take place, is central to the common-sense concept of a cause" (Hart & Honoré 1985, p. 29). They argue that our judgments about what constitutes the normal course of events are guided context-sensitively – sometimes by what usually happens, and sometimes by social, moral, and legal norms. Their account readily explains why we regard Professor Smith's action rather than the administrative assistant's as the cause of the problem: for his action makes a difference to what happens *normally* – that is, in conformity with the prevailing norms – in a way that the administrative assistant's does not.

Hart and Honoré's account of the way our causal judgments are shaped by moral considerations is better than Hitchcock and Knobe's for several reasons: (1) Hart and Honoré's account captures in a seamless fashion the idea that causes are difference-makers for their effects. In contrast, it isn't clear how Hitchcock and Knobe's account captures this idea. Is it through the link with counterfactuals or through the rules about counterfactual availability? (2) Hart and Honoré's account doesn't tie people's causal judgments so closely with their counterfactual judgments, which is a virtue given the empirical evidence dissociating them. If it makes a link with counterfactuals, it is with counterfactuals that are based not on the actual world but on "normalised" worlds that abstract from the abnormal features of the actual world (Menzies 2007). (3) Hart and Honoré's account provides a more uniform account of the contrastive structure of actual causation. Many philosophers have observed that causal judgments have an implicit contrastive structure: the causal judgment "*c* caused *e*" has the implicit contrastive structure "*c* rather than *c** caused *e* rather than *e**." People typically select as the contrast elements *c** and *e** events that would normally have occurred if the abnormal actual events *c* and *e* had not occurred (Menzies 2009). This follows straightforwardly from Hart and Honoré's account, which incorporates the contrastive structure into the semantic content of causal judgments. If Hitchcock and Knobe's account is to explain the contrastive character of causal judgments, it must do so through appealing to pragmatic or non-semantic rules about counterfactual availability.

Neither moralists, nor scientists: We are counterfactually reasoning animals

doi:10.1017/S0140525X10001846

Bence Nanay

Department of Philosophy, University of Antwerp, 2000 Antwerp, Belgium; and University of Cambridge, Cambridge CB2 1RD, United Kingdom.

bence.nanay@ua.ac.be **http://webh01.ua.ac.be/bence.nanay**

bn206@cam.ac.uk

Abstract: We are neither scientists nor moralists. Our mental capacities (such as attributing intentionality) are neither akin to the scientist's exact reasoning, nor are they "suffused through and through with moral considerations" (Knobe's target article, sect. 2.2, last para.). They are more similar to all those simple capacities that humans and animals are equally capable of, but with enhanced sensitivity to counterfactual situations: *of what could have been*.

Knobe presents us with a false dilemma on the level of the metaphors he uses: maybe we are neither scientists nor moralists. But he also presents us with a false dilemma when it comes to the two explanatory schemes he considers: The first one is that the

competences that underlie our mental capacities (to attribute intentionality or to spot causal relevance) are influenced by moral considerations. The second is that these competences are themselves non-moral, but there is some additional factor that makes it the case that our attribution of intentionality is influenced by moral considerations. I will focus on the attribution of intentionality that Knobe considers to be the strongest case in favor of his claims.

The two options Knobe offers are not exhaustive. In fact, they share a premise that we have good reasons to doubt: that is, the premise that the attribution of intentionality is influenced by moral considerations. Knobe's reason for holding this claim is that in two very similar scenarios, the "harm" and the "help" scenarios (Knobe 2003a; see also sect. 3.1 of the target article) that differ only in their moral overtones, our attribution of intentionality also differs. As he says, "the only major difference between the two vignettes lies in the moral status of the chairman's behavior" (sect. 3.1, para. 2).

But that is definitely not the only major difference (see Nanay [2010] for an overview). One striking feature of the experiments Knobe and his collaborators conducted on this topic is that they all share the same structure. To put it very simply, in one scenario, the agent has two reasons for performing a certain action and ignores one of these; in the other, the agent has a reason for and a reason against performing an action and ignores the reason against. Thus, in Knobe's most famous helping/harming experiment (Knobe 2003a; see also target article, sect. 3.1), we have the following two scenarios:

(a) In the harm case, the chairman has a reason (R_1) for introducing the plan (i.e., to increase profit) and a reason (R_2) against (i.e., to avoid harming the environment).

(b) In the help case, in contrast, the chairman has two different reasons to introduce the plan: he had a reason to increase the company's profit (R_1) and he also had a reason to help the environment (R_3).

In short, the difference between scenario (a) and scenario (b) is that in (a) the chairman has R_1 for and R_2 against introducing the plan, whereas in (b) he has R_1 and R_3 both in favor of performing this action. Importantly, the chairman chooses to *ignore* the environmental considerations: R_2 and R_3, respectively. This leaves R_1 in both scenarios, which is a reason for introducing the plan. There is no difference between (a) and (b) in the actual reason the chairman is acting on.

But there is a modal difference between (a) and (b): a difference in what would happen if the chairman *did not ignore* R_2 and R_3, respectively. Contrast the original scenarios (a) and (b) with another pair of cases where the chairman chooses *not* to ignore the environmental considerations:

(a*) The chairman chooses not to ignore R_2 (i.e., a reason against introducing the plan). Then his action would, or at least it could, be different, as now he has a reason for (R_1) and a reason against (R_2) introducing the plan.

(b*) The chairman chooses not to ignore R_3 (i.e., a reason for introducing the plan). His action *would still be the same*, as now he has two reasons (R_1 and R_3) in favor of introducing the plan.

So an important difference between case (a) and case (b) is a modal one: The outcome would be different if the chairman didn't ignore the environmental considerations. In (b), ignoring that the plan helps the environment would make no difference, as there are two independent reasons in favor of introducing the plan: The chairman's actions in (b) and in (b*) will be the same. In (a), on the other hand, ignoring that the plan harms the environment would make (or at least it could make) a difference: The chairman's actions in (a) and in (a*) will be (or at least can be) different.

Thus, what this experiment shows is that in (b), introducing the new scheme does not depend counterfactually on ignoring the environmental considerations, whereas in (a), there is counterfactual dependence between ignoring the environmental considerations and introducing the new scheme. This counterfactual dependence in (a) is not very strong, as not ignoring will not guarantee that the chairman's action will be different, but it is an instance of counterfactual dependence nonetheless. In (b), we have no counterfactual dependence, weak or strong.

What I have said so far shows that the experimental data Knobe uses can be explained with the help of an alternative hypothesis, where the attribution of intentionality does not depend on our moral judgments. In other words, we have two ways of explaining Knobe's original experiments: one appeals to moral judgments, the other one does not. The fact that my explanatory scheme is consistent with Knobe's findings in itself casts doubt on his conclusion.

But we can say something even stronger. My explanatory scheme is in fact preferable to Knobe's for two reasons. First, my explanatory scheme is more robust than Knobe's: it can explain cases of the attribution of intentionality that Knobe's cannot. There are several scenarios where we get differences in the attribution of intentionality without any moral difference (Machery 2008; Mallon 2008; Nanay 2010; Nichols & Ulatowski 2007; maybe even Knobe 2007). As these cases all follow the modal asymmetry I identified, I can account for them (Nanay 2010). Knobe cannot.

Second, those of us with naturalist leanings prefer to explain our complex mental capacities in simple terms. When explaining the mental capacity of attributing intentionality to others, the (broadly) naturalistic way to proceed would be to account for this mental capacity with reference to simple mental processes. This is exactly my strategy: If we can explain the attribution of intentionality with reference to mental capacities that nonhuman animals also possess, plus some further ability to be sensitive to counterfactual situations (which at least some nonhuman primates may also possess; see Suddendorf & Whiten 2001), we should not rely on any further, uniquely human higher-order phenomena, such as morality.

Ambiguity of "intention"

doi:10.1017/S0140525X10001858

Thomas M. Scanlon

Department of Philosophy, Harvard University, Cambridge, MA 02138.
scanlon@fas.harvard.edu

Abstract: Knobe reports that subjects' judgments of whether an agent did something intentionally vary depending on whether the outcome in question was seen by them as good or as bad. He concludes that subjects' moral views affect their judgments about intentional action. This conclusion appears to follow only if different meanings of "intention" are overlooked.

Knobe describes a number of studies in which, he claims, subjects' moral judgments influence their views about whether the actions of others were intentional, about whether an agent did something or merely allowed it to happen, and about whether an agent caused an undesirable consequence. He concludes that the exercise of competencies that humans use in making what might seem to be purely factual judgments about the world – such as judgments about causes and judgments about other agents' mental states – is "suffused with moral considerations from the very beginning" (sect. 5.3, para. 3).

Knobe suggests, very plausibly, that people's judgments about "the cause" of an event depend on a selection of relevant alternatives. His experimental evidence supports the conclusion that in some cases moral considerations partly determine this selection, although it remains an open question how wide this range of cases is. The same may well be true of judgments distinguishing between "doing" and "allowing."

In this comment, however, I will focus on Knobe's claims regarding judgments about intentional action. Here his

conclusions seem to me not to be supported by the evidence he describes, because there is an alternative interpretation of his experimental results that is more plausible than the one he proposes.

The use of *intentional* and its cognates involves a well-known ambiguity (see Anscombe 1958, p. 9; Scanlon 2008, p. 10). One sense of "intentional" is the one opposed to "unintentional." An agent does something "intentionally" in this sense if he or she realizes that this is what he or she is doing – call an action that is intentional in this sense *belief-intentional*. An agent's intention in the other sense is what he or she aims at in so acting. What an agent does "intentionally" in this other sense is opposed to what he or she sees as a mere side-effect of so acting – call what is intentional in this sense *aim-intentional*.

The effects on the environment of the policies adopted by the chairmen in the two experiments Knobe describes are belief-intentional: the description of the cases makes clear that they are aware that the policies they choose will have these effects. But these effects are not aim-intentional: the descriptions make clear that the chairmen are indifferent to these effects, and are concerned only with profits. Given that these facts are made clear in the presentation of the scenarios, it is reasonable to believe that the subjects in each case have the same beliefs about the chairman's mental state: that the bringing about of these effects is belief-intentional but not aim-intentional. The differing answers that the subjects give to the question of whether the chairmen harmed or helped the environment "intentionally" is indeed due to moral considerations, but not in the way that Knobe suggests.

The important moral fact here is that agents are commonly open to moral criticism for bringing about bad effects when they know that these effects will occur even if they do not aim at these effects – that is to say, when they do so belief-intentionally, even if not aim-intentionally. But agents are generally held to merit moral praise or credit for bringing about good consequences only if they do so aim-intentionally. Given that the subjects see harm to the environment as a bad thing, when they are asked whether the chairman in the first scenario harmed the environment intentionally, what they are likely to ask themselves is whether the chairman's action was intentional in the sense relevant to moral criticism for bringing about such an effect (that is to say, whether it was belief-intentional). In the other case, since the subjects are likely to view helping the environment as a good thing, when they are asked whether the chairman helped the environment intentionally, what they are likely to ask themselves is whether what the chairman did was intentional in the sense relevant to moral praise or credit (that is to say, whether it was aim-intentional). What the shift from harming to helping does is not to change the subjects' interpretation of the chairman's mental states in the respective scenarios, but rather, to change the question about those mental states to one that seems to the subjects to be relevant.

This interpretation of the subjects' responses seems to me extremely plausible. It is also supported by some of the further details that Knobe mentions. For example, he reports that when subjects are asked whether "the chairman intended to harm the environment," answers are moved strongly in a negative direction (sect. 3.2, para. 4). This is to be expected on the interpretation I propose, because the verb *intend* suggests (aim) intention more strongly than does the adverb *intentionally*.

My interpretation also explains why subjects disagree with the claim that an agent was "in favor of" a morally good outcome but are neutral on the question of whether the agent was "in favor of" a morally bad outcome (sect. 3.2, para. 5). This is because the agent fails to favor the morally good outcome *in the way relevant to moral praise or credit*; but, even if he or she does not actively favor the morally bad outcome, an agent who is perfectly willing to bring about that outcome for some other reason is more favorably disposed toward it than he or she should be, and therefore open to some criticism on this score.

Alternatives and defaults: Knobe's two explanations of how moral judgments influence intuitions about intentionality and causation

doi:10.1017/S0140525X1000186X

Walter Sinnott-Armstrong

Philosophy Department and Kenan Institute for Ethics, Duke University, Durham, NC 27708.

ws66@duke.edu

http://kenan.ethics.duke.edu/people/faculty/walter-sinnott-armstrong/

Abstract: Knobe cites both relevant alternatives and defaults on a continuum to explain how moral judgments influence intuitions about certain apparently non-moral notions. I ask (1) how these two accounts are related, (2) whether they exclude or supplement supposedly competing theories, and (3) how to get positive evidence that people consider relevant alternatives when applying such notions.

Joshua Knobe's novel theory of how moral judgments influence people's intuitions about certain apparently non-moral notions, including intentionality and causation, is a version of contrastivism (see Sinnott-Armstrong 2008). As with many other topics (including knowledge, free will, explanation, and morality), it is illuminating to consider the range of possible contrasts or alternatives and ask when and why people limit their attention to a smaller contrast class. So I am very sympathetic. I would, however, like to press Knobe to develop three aspects of his theory.

First, Knobe formulates his "general approach" in his section 5.1 in terms of relevant alternatives. Moral judgments are said to affect which counterfactual alternatives are seen or treated as relevant. Next, Knobe discusses his "case study" in his section 5.2 in terms of defaults on a continuum. Moral judgments are said to affect the position of the default. These views are not equivalent, because alternatives need not always fall on a continuum, and relevant alternatives might fall on either side of a default. Knobe describes the default as "a particular sort of alternative possibility," but it does not seem to be the only relevant alternative, so comparing a default and comparing a range of alternative possibilities seem quite different. My question for Knobe is then: What exactly is the relation between these two theories?

Second, it is also not clear what the relation is between either of Knobe's suggestions and the views against which he has argued in the earlier parts of his article. Motivation, blame, emotion, and pragmatic context would seem to be promising candidates for explaining why we treat certain alternatives rather than others as relevant, or why we place the default at one point instead of another on a continuum. If Knobe agrees, then his own theory, though a crucial part of the story, would need to be supplemented by central aspects of the views he criticizes. His theory then works together with his supposed opponents, rather than supplanting them. But if Knobe denies that these features explain why we adopt certain relevant alternatives and defaults, then we need an alternative explanation of relevance and default. It is not enough to refer to alternatives and defaults without explaining how the alternatives and defaults get set.

Addressing this issue, Knobe says, "all sorts of different factors can play a role here" (sect. 5.1, para. 5). This is surely right, and he cites supporting literature. However, it leaves open the possibility that motivation, blame, emotions, and pragmatic context do sometimes play roles in determining which alternatives we see as relevant and where we place the default. No theory that focuses on one single factor can or should be expected to cover all examples, even if each factor does explain some variance in some areas. Hence, I also want to ask Knobe whether his arguments are supposed to show that motivation, blame, emotions,

and pragmatic context do not *always* play a role or do not *ever* play a role in determining relevant alternatives and defaults.

Third, after he criticizes his opponents for failing "to produce any positive evidence in favor of the hypothesis" (sect. 4.1.4, para. 1), I would like to see Knobe's positive evidence in favor of his own hypotheses. Consider first his claim about relevant alternatives, and focus on his example of what caused the dent in the car. He suggests that, when we ask whether a certain person, object, or event caused a certain effect, we "think about," "consider," "compare," and "pick out just certain specific alternatives" (sect. 5.1, para. 4). As Knobe says, we do not "consider the possibility that the car could have been levitating in the air." But what is the positive evidence that we do consider other alternatives?

I do not deny that we treat a range of alternatives as relevant. However, it is not clear whether we actually represent these alternatives, even unconsciously. Another possibility is that we have a disposition to dismiss certain alternatives as irrelevant, if raised, and to accept other alternatives as relevant, if raised; but we never explicitly "think about" or "consider" the relevant alternatives any more than the irrelevant alternatives unless prompted.

How can we decide between these views? Perhaps we could get evidence that subjects consider or think about certain alternatives by asking the subjects, but self-report would not be reliable. Another method would be to measure subjects' memory errors, word completion patterns, or reaction times when asked whether certain words were in the scenarios. If subjects really do consider an alternative that would naturally be formulated in certain terms that were not actually in the scenario, then we would expect them to be more likely to misremember those terms as being in the scenario, to complete letter strings so as to form those terms, and to unscramble the letters from those terms more quickly than if they never considered that alternative.

However, before we can apply these techniques, we would need to formulate specific hypotheses about which alternatives are and are not thought about or considered in which scenarios. It is not enough simply to say that moral judgments affect the range of alternatives that are taken to be relevant. We need to know which alternatives are supposed to be seen as relevant. Only then can we test whether those alternatives are actually considered or thought about, as Knobe claims.

The same basic issue arises for Knobe's theory that moral judgments affect which point on a continuum is seen as the default. Certain hypotheses might seem plausible and might have explanatory power, but it is not easy to figure out how to gather positive evidence for the hypothesis that people actually set different defaults depending on their moral judgments. So my last question for Knobe is: How will you get positive evidence for your claims that moral judgments affect relevant alternatives and defaults?

"Very like a whale": Analogies about the mind need salient similarity to convey information

doi:10.1017/S0140525X10001871

David Spurrett and Jeffrey Martin

School of Philosophy and Ethics, University of KwaZulu-Natal, Durban 4041, South Africa.

spurrett@ukzn.ac.za **drcogsci@gmail.com**
http://ukzn.academia.edu/DavidSpurrett
http://ukzn.academia.edu/JeffMartin

Abstract: Knobe relies on unhelpful analogies in stating his main thesis about the mind. It isn't clear what saying the mind works, or doesn't work, "like a modern university" or "a scientific investigation" means. We suggest he should say that some think that human cognition respects a ban on fallacies of relevance, where considerations actually irrelevant to truth are taken as evidence. His research shows that no such ban is respected.

We have no dispute with Knobe's description of the ways in which various accounts – including motivational bias and conversational pragmatics – of the experimental results he describes don't succeed, and how the competency approach he favours currently does better.

This said, it remains unclear precisely what Knobe's position is, because his exposition depends on analogies that are both underdeveloped and problematic. Consequently, the answers to two questions are not sufficiently clear. The first question is: What specific commitment regarding human cognition is being *rejected*? The second question is: What specific claim about human cognition is being *defended*?

Analogical reasoning transfers information from one object to another, non-identical one. For this to work well, the two objects need to have enough salient features or relationships in common. Disanalogies between paradigmatic features of the two objects impede the transfer. Niels Bohr, for example, explained his rejection of previous models of the atom partly by drawing an analogy with the solar system. Despite important differences between atoms and planetary systems, this was a good way of getting at a few key and, at the time, radical ideas: Atoms are mostly empty; very small parts of them are in approximately orbital activity around other central ones.

Knobe offers two analogies for the view he is ostensibly rejecting. According to the first, the human mind works "something like a modern university" (sect. 1, para. 2). According to the second, which is an analogy within the first, some mental processes use the "same sorts of methods we find in university science departments" (sect. 1, para. 2). Except for relatively cryptic remarks on the ways disciplines are supposedly separated in universities, and some (also brief) remarks on science, Knobe develops neither analogy in significant detail.

What might it mean for the mind to be like a university? Knobe suggests that the organisation of a university corresponds to a set of distinctions between types of questions, so that the mind has something analogous to theology, art, philosophy, and some scientific departments. He goes on to argue that the mind is not like this. But the administrative organisation of universities into departments exists along with a patchwork of overlapping techniques, theories, problems, and collaborative research programmes cutting across departmental divisions. The analogy also doesn't do the work Knobe requires because some departments, such as those of history and politics, consider both factual and moral questions, just as art departments consider factual and aesthetic ones (not merely "is this painting good?" but also "is it genuine?"). Philosophy departments notoriously consider almost anything – these days they even do experiments.

The fact that the overall organisation of universities is not consistently or strictly modular need not be a big problem, since most of the heavy lifting is done by the second analogy, suggesting a view (the one to be rejected) where some mental processes use the same methods as scientists do. Unfortunately, though, there are no agreed upon set of criteria that separate science from non-science, partly because there is no clear division between the methods of "science" and those of other enterprises. Philosophers of science have argued for generations without converging on consensus about what, if anything, demarcates science from pseudo-science and non-science. That this is so is reason to recognise that "like science" is not a promising explanatory analogy.

We suggest that neither analogy need be repaired, or even replaced. Instead, the claim at issue can be stated directly. Knobe gives us a clue when he says that "Genuinely scientific inquiry seems to be sensitive to a quite specific range of considerations and seems to take those considerations into account in a highly distinctive manner" (sect. 2.1, para. 5). We think it makes most sense to read this as saying that the "specific range of considerations" are *epistemic* considerations, which is to say ones strictly relevant to whether or not some claim is true. It is a good normative rule for truth seekers to avoid fallacies of

relevance. One example of such a fallacy is an appeal to consequences. Saying that evolution by natural selection should be rejected because believing it (supposedly) leads to selfishness, appeals to considerations which have no evidential value. Likewise, that an experimenter is very nice, or nasty, or eccentric, has no epistemic value as far as the empirical test of a hypothesis itself is concerned.

The phenomena to which Knobe draws our attention, and which his own empirical work has done a great deal to document, are all examples of fallacies of relevance, mostly in the attribution of credit for intention and causation. Whether someone caused something, intended it, or is responsible for it, depends on what they did and how that influenced the world. It does not depend on whether what happened is the sort of thing we would regard as morally objectionable. The fact that considerations relating to the moral value of the *outcomes* appear to affect judgements regarding what was *intended*, or *caused*, suggests that some of our mental processes are routinely prone to what, by responsible epistemic lights, are fallacies of relevance.

The general claim about human cognition that Knobe is rejecting, we therefore suggest, is one to the effect that the organisation of (human) cognition respects this normative standard, and that it does so by not allowing strictly irrelevant considerations to interact during processing. We already have ample evidence that the general claim is false, from, among other things, a long history of social psychology and behavioural economic experiments. Thorndike (1920), for example, showed that in assessments of other people, perceptions of some traits were more correlated with perceptions of other traits than should be the case if traits (such as attractiveness and competence) varied independently. What is exciting and surprising about the work Knobe reviews (and has been conducting himself) is that, from this point of view, it shows the persistent influence of moral reactions in judgements about matters where those reactions are irrelevant to truth.

It would be interesting, not to mention extremely important, to see whether the effects are reduced when people deliberate about causation and responsibility in organised groups charged with an epistemic task – for example, juries.

Are we really moralizing creatures through and through?

doi:10.1017/S0140525X10001883

Stephen Stich[a] and Tomasz Wysocki[b]

[a]*Department of Philosophy, Rutgers University, New Brunswick, NJ 08901-1107;* [b]*Institute of Philosophy, University of Wroclaw, ul. Koszarowa 3, 51-149 Wrocław, Poland.*

sstich@ruccs.rutgers.edu **http://www.rci.rutgers.edu/~stich/**
tomaszwysocki@xphi-europe.org **http://www.xphi-europe.org/**

Abstract: Knobe contends that in making judgments about a wide range of matters, moral considerations and scientific considerations are "jumbled together" and thus that "we are moralizing creatures through and through." We argue that his own account of the mechanism underlying these judgments does not support this radical conclusion.

In his conclusion, Knobe reminds us that the target article began with a metaphor (well, a simile, actually) comparing the organization of the mind to the organization of a 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" (sect. 6, para. 1) This suggests a view on which moral judgments play a quite limited role in cognition.

> 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 (sect. 6, para. 2).

Knobe maintains that this picture is "deeply mistaken":

> [There is no] 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*. (sect. 6, para. 3, emphasis added)

This is a bold and radical view. And while we share Knobe's fondness for views that fly in the face of conventional wisdom, we are not persuaded that he has made a convincing case. Indeed, we think that Knobe's own explanation for the sorts of phenomena he so clearly and carefully documents flies in the face of these audacious and dramatic claims.

To explain our skepticism, we will focus on the target article's Figures 6–8. These are aimed at explaining how people make judgments about whether an agent is *in favor* of an outcome. Knobe begins his explanation with the "fundamental assumption" 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. . . . For simplicity, we can depict this continuum in terms of a scale running from *con* to *pro*. (Fig. 6) (sect. 5.2, para. 2)

An agent whose attitude falls way over on the *con* side, Knobe tells us, will be classified as "not in favor" and an agent whose attitude falls way over on the *pro* side will be classified as "in favor." But that does not tell us "how . . . people determine the threshold at which an agent's attitude passes over from the category 'not in favor' to the category 'in favor'" (sect. 5.2, para. 2). To explain this, Knobe posits "an additional element" that includes a variable default position whose location along the continuum is determined, in part, by people's moral judgments. Knobe proceeds to tell us, in some detail, how this default-setting system works, and how it plays a role in determining whether we judge that an agent is in favor of the outcome in question.

There is, however, one central and important part of the system about which Knobe tells us nothing at all. The lacunae emerges very clearly when we compare Figure 7 to Figure 8. One difference between these two figures, the one that Knobe focuses on, is that the Default position, and thus the part of the continuum that supports a judgment that the agent is "IN FAVOR," has been shifted to the right. But there is another difference. In Figure 8, the position of the Agent on the continuum has been marked. And that position is, of course, crucial to the account. In Figure 8, the Agent is located to the left of the Default, leading to a judgment that the Agent is not in favor of the outcome in question. Had the Agent been located significantly further to the right, the system would produce the judgment that the Agent *is* in favor of the outcome.

But how does the psychological mechanism that Knobe posits succeed in locating the Agent along the continuum? As far as we can see, Knobe tells us nothing about this, and there is certainly no hint that the psychological processes responsible for locating the Agent along the continuum are sensitive to any moral or evaluative judgment made elsewhere in the system. Rather, it seems, this crucial determination is made in a value-free way. To revert to Knobe's recurrent metaphor, it is made by one of the mind's "science departments" that focuses on "developing a purely 'scientific' understanding of what is going on in a situation and remain[s] neutral on all questions of morality" (sect. 6, para. 2).

The point is underscored by Knobe's analogy with the process that a teacher might use in assigning grades. The teacher starts out with a "continuum of different percentage scores on a test" (sect. 5.2, para. 4) and must then decide on a threshold beyond which a score will count as an A. Her process for setting the

threshold involves a variable default determined by the teacher's assessment of the difficulty of the test. This is analogous to the variable, morally influenced default depicted in Figures 7 and 8. However, presumably the test scores themselves are *not* influenced by the teacher's assessment of the difficulty of the test. They, like the position of the Agent in Figure 8, are determined by a "purely scientific" component in the assessment process.

The bottom line is that on Knobe's own account of how we decide whether an agent is in favor of an outcome, there *is* a clear division between psychological processes that involve moral considerations and those that do not. It is not the case that "everything is jumbled together," nor is it the case that "we are moralizing creatures through and through."

Depression affecting moral judgment

doi:10.1017/S0140525X10001895

Luisa Terroni[a] and Renerio Fraguas[b]

[a]*Liaison Psychiatry Group, Department of Psychiatry, Institute of Psychiatry, Clinics Hospital, Medical Faculty, University of São Paulo, São Paulo, Brazil;* [b]*Liaison Psychiatry Group, Laboratory of Psychiatric Neuroimaging (LIM-21), Department of Psychiatry, Institute of Psychiatry, Clinics Hospital, Medical Faculty, University of São Paulo, São Paulo, Brazil.*

luterroni@gmail.com **http://www.fm.usp.br**
fraguasr@gmail.com **http://www.fm.usp.br**

Abstract: Depressive mood can be involved in the moral judgments made by people with depression. Here, we focus on the negative judgments depressed patients have of themselves and the world. Possibly, the alterations in moral judgment in subjects with depression can be understood by taking into account the neural basis of depression.

In his article, Knobe discusses the role of moral judgments in people's understanding. The author focuses his study on the moral influence present in the process of cognition. In discussing the motivational bias hypothesis (sect. 4.1), Knobe mentions studies in patients who had a lesion in the ventromedial prefrontal cortex. These studies tried to demonstrate the non-involvement of affective reaction in the process of moral judgment.

In this commentary, we consider moral judgment in depressed subjects. We feel that depressive mood is particularly relevant to the negative moral judgments often made by patients with depression. In the psychopathology of depression, patients' actions and thoughts can be affected by the depressive mood, which in turn tends to affect their moral judgments. This psychopathological process is characterized by ruminations of negative thoughts. Patients with major depression understand the world and themselves in the same way, that is, in a negative way. Here, these disturbed thoughts can be understood as a distortion of moral judgments influenced by the presence of depressive mood. In the moral judgments, patients often evaluate themselves or their acts as something "bad" or "wrong." This process supposes that patients are evaluating themselves axiologically. During this process, depressed patients consider themselves to "blame." Such depressed patients with these negative moral judgments and thoughts can be an example for the line of study developed by Knobe. Patients' negative moral judgment and other psychopathological alterations return to normal with the remission of depression (Benedetti et al. 2007). This process of continuous and rigid negative moral judgments may have a biological explanation.

What kind of process can be underlying this alteration on negative moral judgment in depression? In normal human subjects, studies with functional brain imaging have found increased activity in brain areas in a resting state and reduced activity when there is a proposed goal for brain function . This organized mode of brain function identified in specific brain regions constitutes the default system (Drevets et al. 2008). A disturbance in this network can explain depression symptomatology. Studies developed by Sheline et al. (2009) found that depressed subjects showed less decrease in activity than control subjects in areas of the default system, or default mode network (DMN), during performance of emotional tasks. These findings have supported the view that alterations in areas of the DMN may constitute a basis for the disordered self-referential thoughts of depression.

In Knobe's article, there is a mention about a study investigating cognition in people who don't have immediate affective reactions as a result of lesions in the ventromedial prefrontal cortex. He uses this study to show that the results of his analyses about moral judgment in normal subjects are not due to an affective reaction. However, in people with distortion of judgment caused by depressive mood, the neural dysfunction in specific brain areas found in depression investigations can explain the moral judgment disturbance, and supports the interference of the depressive mood on moral judgments.

This construction supports our view that the neural basis of depression may explain depressive mood and, consequently, moral judgment; albeit it does not exclude Knobe's point that moral judgment may occur independently of affective reaction.

Fixing the default position in Knobe's competence model

doi:10.1017/S0140525X10001901

Joseph Ulatowski[a] and Justus Johnson[b]

[a]*Department of Philosophy, University of Nevada–Las Vegas, Las Vegas, NV 89154-5028;* [b]*Department of Philosophy, University of Wyoming, Laramie, WY 82071.*

oohlah@unlv.edu **justus.johnson@me.com**
http://web.mac.com/oohlah

Abstract: Although we agree with the spirit of Knobe's competence model, our aim in this commentary is to argue that the default position should be made more precise. Our quibble with Knobe's model is that we find it hard to ascribe a coherent view to some experimental subjects if the default position is not clearly defined.

In the target article "Person as Scientist, Person as Moralist," Joshua Knobe has devised an innovative model where moral appraisals play a fundamental role in how people make sense of agents and their actions. According to Knobe, people's intuitions depend on a comparison between the action under consideration and an alternative possibility, which he calls the "default position." The default position falls somewhere along a continuum, but experimenters fail to designate its exact location. In this commentary, we contend that the default position must be fixed and clearly articulated. Otherwise, some of the subjects' intuitions seem incoherent. We agree with Knobe that there seems to be a default position against which people judge whether or not some action under consideration is favored. But we believe that his approach may be made more precise than it is by specifying clearly what the default position is.

According to Knobe's competence model, moral considerations figure into how subjects make a comparison between the action under consideration and certain alternative possibilities. It seems people who view an action as morally bad uphold an attitude at least slightly toward the *con* side, and people who view an action as morally good tend to have an attitude at least slightly toward the *pro* side. An action is favored when "the agent's attitude falls sufficiently far beyond the default" (sect. 5.2, para. 5). The *core* of Knobe's explanation has it that "moral judgments affect [people's] intuitions *by shifting the position of the default*" (sect. 5.2, para. 6, emphasis Knobe's).

Knobe's competence model has done a nice job of explaining why a majority of subjects answered the harm and help scenarios

as they did (sect. 3.1). When a majority of subjects (82% according to Knobe 2003a) compare the chairman's attitude to the default position that harming the environment is morally bad, they favor the response that the chairman intentionally harmed the environment. Likewise, when a majority of subjects (77% according to Knobe 2003a) compare the chairman's attitude to the default position that helping the environment is morally good, they do not favor the response that the chairman intentionally helped the environment. Although Knobe's competence model has succeeded in explaining the majority's intuitions, his model may not succeed in explaining the intuitions of subjects who gave the minority view.

Some subjects responded that the chairman did not intentionally harm the environment (18%) or that the chairman did intentionally help the environment (23%) (Knobe 2003a). These represent a *minority response* in the harm case and help case, respectively. If Knobe's competence model is correct, then the minority's default position for the harm scenario is that harming the environment is a morally good thing. The data also suggest that the minority's default position in the help case is that helping the environment is a morally bad thing. These views are unusual and the result of applying Knobe's competence model.

Knobe may object to this assessment. Subjects receiving the harm scenario may hold that harming the environment is morally bad but the chairman's indifference does not constitute that he intentionally harmed the environment. These subjects may refrain from saying that the chairman intentionally harmed the environment because the chairman did not want to harm the environment. Subjects who responded that the chairman intentionally helped the environment may uphold the default position that helping the environment is morally good. Since the chairman knew that the program would help the environment, subjects chose the response that he intentionally helped the environment. If this is correct, Knobe is able to show why the competence model explains the minority's intuitions.

The problem with this response is that one of us (Ulatowski) collected data where two-thirds of subjects given both the harm and the help case chose minority responses (Nichols & Ulatowski 2007). Of the two-thirds, half responded that:

> (1) The chairman intentionally helped the environment and the chairman intentionally harmed the environment.
>
> or that:
>
> (2) The chairman did not intentionally harm the environment and the chairman did not intentionally help the environment.

On response (1), if the competence model is correct, then respondents judged that not only is harming the environment morally bad but helping the environment is morally bad, too. On response (2), if the competence model is correct, the default position for subjects is not only that helping the environment is morally good but that harming the environment is morally good, too. Subjects' default positions seem to be inconsistent and, therefore, incoherent.

We suggest that the default position be clearly defined to avoid incoherence of subjects' intuitions. In a series of experiments testing whether the distinction between *doing* and *allowing* depends on moral appraisals, we specified an alternative possibility against which subjects should compare the agent's action (Ulatowski & Johnson 2010):

> Five people are in imminent danger of death, and you are a part of a team that is taking a special train to rescue the five people. Every second counts. You have just taken over from the driver, who has gone to the back of the train to check on something. Since the train is on automatic control, you don't need to do anything to keep it going. But you can stop it by putting on the brakes. You suddenly see someone trapped ahead on the track. If you don't do anything, he will be killed (though the train will be able to continue on its way). But if you do stop, and then free the man, the rescue mission will be aborted. So you let the train continue.

We asked subjects, "Since you could have stopped the train, did you kill the man on the track?" We stipulated the default position: to stop the train. We believe that by specifying the default position, it may prevent an incoherent interpretation of people's intuitions.

Our aim in this commentary has been to expose the incoherence in subjects' responses when an experiment fails to stipulate the default position. Ultimately, we cannot assume that we know what the subjects' default position is.

ACKNOWLEDGEMENTS

We would like to thank Dave Beisecker for helpful conversation leading to a draft of this commentary, and Elijah Millgram for comments on an earlier draft.

Author's Response

The person as moralist account and its alternatives

doi:10.1017/S0140525X1000230X

Joshua Knobe

Program in Cognitive Science and Department of Philosophy, Yale University, New Haven, CT 06520-8306.

joshua.knobe@yale.edu **http://pantheon.yale.edu/~jk762/**

Abstract: The commentators offer helpful suggestions at three levels: (1) explanations for the particular effects discussed in the target article; (2) implications of those effects for our understanding of the role of moral judgment in human cognition; and (3) more theoretical questions about the overall relationship between ordinary cognition and systematic science. The present response takes up these three issues in turn.

The commentators have offered helpful suggestions and criticisms at all levels, from the nitty-gritty of the individual experiments to the broadest sorts of theoretical and philosophical issues. Clearly, the questions at these different levels are intimately connected, but since one has to begin somewhere, perhaps it is best to start by focusing in on the trees and then move gradually toward thinking about the overall shape of the forest. In other words, we can start with specific questions about the explanations for particular effects and then move to implications for broader theoretical and philosophical issues.

R1. Alternative hypotheses

Recent studies indicate that people's moral judgments can impact their application of a surprising range of different concepts. Moral judgments appear to be impacting people's application of the concepts of intentional action, causation, freedom, knowledge, doing and allowing, desire, and many other concepts besides. The primary aim of the target article was to provide an explanation for this pervasive impact of moral judgment.

To explain these phenomena, I offered a specific hypothesis. The suggestion was that people come to an understanding of the actual world by comparing it with certain alternative possibilities (counterfactuals). People's

moral judgments impact their selection of alternatives and thereby influence their application of a wide range of different concepts.

A number of commentators responded by developing competing hypotheses. These hypotheses explain the impact of moral considerations in terms of quite different sorts of cognitive processes.

R1.1. A case study

One worry about many of these hypotheses is that they proceed by picking out just one concept whose application is affected by moral judgment and examining this one concept in isolation from all the others. Hence, these hypotheses offer explanations for one of the effects of moral judgment but say nothing about other effects that seem, at least initially, to be closely related.

Of course, the fact that a hypothesis is framed entirely in terms of one of these effects does not mean that this hypothesis has to be incorrect. Future research might show that the hypothesis can be extended in fairly natural ways to handle other related phenomena, or perhaps it will be shown that the phenomena that initially seem so closely related are, in fact, fundamentally different. The problem, then, is not that these hypotheses are necessarily wrong but just that they have not yet been developed to the point where they can be properly evaluated.

Thus, to take one example, **Scanlon** suggests that we might be able to explain the apparent asymmetries in people's intuitions about intentional action by looking more closely at the meaning of the word *intentionally*. Specifically, suppose we assume that an expression like "John brought about the outcome intentionally" actually has two distinct meanings:

(a) John knew that he was bringing about the outcome.
(b) John aimed at bringing about the outcome.

People's moral judgments might then impact their intuitions simply by affecting their sense of which of these two meanings is the relevant one in the context at hand.

This hypothesis does seem to do a nice job of accounting for the asymmetries observed in people's intuitions about intentional action, but the first thing to notice here is that the very same effect can be observed for numerous other concepts. When people determine that a foreseen side-effect is morally bad, they are not only more inclined to say that the agent brought it about *intentionally*; they are also more willing to say that she was *in favor* of it, that she *decided* to bring it about, even that she *advocated* it. Presumably, it is not merely a coincidence that we find this exact same effect arising in the application of so many different concepts. So what we really need here is an explanation for the pattern as a whole.

One option would be to extend **Scanlon**'s hypothesis by claiming that the ambiguity posited for the word *intentionally* can also be found in numerous other expressions. For example, one might say that an expression of the form "John advocated the outcome" also has two distinct meanings. Roughly:

(a) John called on people to adopt a policy with the aim of bringing about the outcome.
(b) John called on people to adopt a policy that he knew would bring about the outcome.

But we would then be offering a hypothesis of a very different type. We would no longer just be pointing to some idiosyncratic feature of the word *intentionally*. Instead, we would be positing a general feature of language that led to a systematic ambiguity within a whole class of expressions. And, of course, the methods used for testing the hypothesis would then have to be correspondingly different. We couldn't proceed just by looking at patterns in people's intuitions about intentional action. We would have to introduce a more general claim about word meanings and then evaluate this claim both by gathering data involving people's use of numerous different expressions and by thinking about the ways in which it fit into larger theories about lexical semantics, polysemy, and so forth.

R1.2. Application to further examples

This very same worry also arises, albeit in somewhat different forms, for a number of the other alternative hypotheses. For example:

Nanay points out that people's judgments about the two intentional action cases differ not only from a *moral* perspective, but also from a *modal* perspective. Specifically, he claims that people who are given the harm case make the judgment:

> If the chairman had not ignored the environmental considerations, he would not have harmed the environment.

but that people who are given the help case do not make the judgment:

> If the chairman had not ignored the environmental considerations, he would not have helped the environment.

Nanay then suggests that this difference in people's modal judgments can lead to a difference in people's intuitions about intentional action. Hence, it might be possible to explain the effect without introducing moral considerations in any way.

Menzies argues that the asymmetries observed for people's causal judgments can be explained if we adopt a theory of causal cognition that emphasizes the role of *normality*. Suppose we assume that people only regard an event as a cause to the extent that this event "intervenes in the normal course of events and makes a difference in the way these develop" (para. 5). Now suppose we further assume that people's ordinary notion of normality is not simply a matter of statistical frequency but also takes into account social, legal, and moral norms. Starting from these two assumptions, we arrive at an interesting and surprising conclusion: If both the behavior of the administrative assistant (a perfectly normal behavior) and the behavior of the professor (a violation of social and moral norms) were necessary for the problem to arise, people will tend to pick out the behavior of the professor and regard it, in particular, as the cause of the problem.

Hindriks suggests that we can come to a better understanding of the intentional action effect by applying the legal distinction between *actus reus* (guilty act) and *mens rea* (guilty mind). He then notes that most research in this domain has focused on the impact of people's judgments of the moral status of the agent's action, with the assumption being that these judgments are somehow influencing people's intuitions about intentional action. By contrast, he suggests that people's intuitions might

https://doi.org/10.1017/S0140525X10001810

actually be affected by a judgment of *mens rea*, that is, a judgment about the status of the agent's mental states. In earlier work, Hindriks has spelled out this claim in terms of the hypothesis that people tend to think that an agent *S* intentionally performed an action *A* to the extent that "An agent *S* ϕs intentionally if *S* intends to ψ, ϕs by ψing, expects to ϕ by ψing, and ψs in spite of the fact that he believes his expected ϕing constitutes a normative reason against ψing" (Hindriks 2008, p. 635).

Humphrey argues that the intentional action effects can be given a straightforward Bayesian interpretation. All one needs to consider is the conditional probabilities people assign in the relevant cases. Thus, suppose we compare (a) the conditional probability that the agent harmed the environment intentionally, given that he implemented the program, and (b) the conditional probability that the agent helped the environment intentionally, given that he implemented the program. If one assigns priors in such a way that (a) is greater than (b), it will follow straightforwardly that people should be more inclined to guess that the agent harmed intentionally than they are to say that the agent helped intentionally.

Brogaard agrees that people's intuitions about intentional action are not purely scientific in nature, but she argues that it would also be a mistake to understand them in terms of the judgments people make about whether actions are morally right or wrong. Instead, she claims, we should understand these intuitions in terms of judgments of *desert*. People make judgments about whether the agent deserves a side-effect, or the blame for it, and these judgments of desert end up influencing their intuitions about whether or not the behavior was performed intentionally.

Lombrozo & Uttich note that people ascribe different attitudes in cases of norm violation from the attitudes they ascribe in cases of more ordinary behavior. If we see that a person has chosen to implement a program that has some entirely innocuous effect, we might assume that this person did not actually care very much about the program either way – maybe he just decided to adopt it without much thought. But now suppose, instead, that we saw a person choosing to implement a program that he knew would harm the environment. Since harming the environment is a norm violation, we might immediately conclude that he must have had some strong interest in adopting this program, and we would therefore be more inclined to attribute to him the kind of *pro*-attitude that would lead us to say that he acted intentionally.

Each of these proposals offers interesting suggestions about a particular concept – and many of these proposals will no doubt lead to important new insights – but all of them seem to leave us with a mystery as to why the impact of moral judgment is so pervasive.

For a particularly promising example, consider the hypothesis that **Menzies** offers about people's causal intuitions. Menzies suggests that causal intuitions can be affected in a complex way by judgments of what might be called "normality." Now, it is an interesting question whether this hypothesis is right or wrong. (As it happens, I think that it is completely correct; Hitchcock & Knobe 2009.) However, the key point is that this hypothesis does not explain why the effect we find for the concept of causation can also be found for so many other concepts. Indeed, there is an important sense in which it does not really *explain* the effect for causal intuitions at all. It simply describes a certain pattern in people's application of this concept, without telling us why the concept works like this and not some other way. So this sort of hypothesis gives us a tantalizing glimpse into the phenomenon at work here, but it seems that we will not really have an adequate account until we can offer a more general theory.

If I may be permitted to speculate, it seems to me that contemporary work on these problems is suffering from the legacy of a certain tradition of conceptual analysis. In early work in that tradition, it was thought that we should proceed by developing for each concept a list of necessary and sufficient conditions. The aim was to provide a separate list of conditions for each concept – one list for the concept of intentional action, one for the concept of causation, and so forth. This tradition has now been widely repudiated. None of the commentators on the present target article attempted to provide lists of necessary and sufficient conditions, and I am sure that most of them would agree that such an approach is unlikely to prove fruitful. Yet, though researchers today are anxious to distance themselves from this program of list-making, I suspect that a certain remnant of that earlier tradition still remains. There are still attempts to go through people's various concepts and provide something like an "analysis" for each of them; it's just that these analyses no longer take the form of necessary and sufficient conditions.

In my view, we should make an even more radical break with the tradition. There is simply no use in developing something like an "analysis of the concept of intentional action" and then, separately, an "analysis of the concept of causation." Instead, we should recognize that people's intuitions about each of these concepts are shaped by a number of distinct psychological processes, and that each of these processes in turn influences intuitions about a number of different concepts. So what we really need is not a separate theory for each of the separate concepts but rather unifying theories of the underlying processes. Such theories might not offer us a comprehensive picture of any one concept, but they will allow us to generate specific testable predictions regarding a whole range of different concepts.

R1.3. Motivation to blame

The contribution from **Alicke & Rose** pursues precisely this strategy. They suggest that the phenomena might be explained in terms of a single underlying psychological process that can affect people's intuitions across a wide variety of different domains. Specifically, they suggest that people sometimes experience a motivation to justify attributions of blame and that this motivation can affect their views about intention, causation, and numerous other issues.

In the target article, I had argued that this sort of process could not explain the effects under discussion here. **Alicke & Rose** reply by reviewing some very impressive data from Alicke's earlier work (Alicke 1992), which they take to provide conclusive evidence that people's judgments actually can be distorted by a motivation to blame.

This commentary definitely raises a number of important issues, but I worry that I was not sufficiently clear in

articulating the nature of the disagreement in the target article itself. The thing to keep in mind is that no one is actually trying to refute the key claim made in Alicke's earlier work. In that earlier work, Alicke provides excellent evidence for the claim that people's intuitions can be distorted by a motivation to blame, and none of the people writing on these issues more recently have been trying to call that claim into question. Rather, the issue is just about whether the theory developed in Alicke's earlier work provides the best explanation for a specific class of effects that have been uncovered in more recent work. Some researchers have argued that it can (Alicke 2008; Nadelhoffer 2006a); others have argued that it cannot (Nichols & Ulatowski 2007; Wright & Bengson 2009; Young et al. 2006).

At this point, I think that Alicke's basic theoretical claims about the importance of a motivation to blame have been established beyond reasonable doubt, and there is no need to provide any further evidence for them. The thing to focus on now is just the detailed structure of these particular effects and whether a motivational explanation can account for them. In the target article, I reviewed some of the experimental evidence for the view that it cannot.

R1.4. Sources of evidence

Sinnott-Armstrong raises more or less this same issue about my own preferred account. The account suggests that people's moral judgments affect their counterfactual reasoning, which in turn plays a role in their application of numerous different concepts. But, Sinnott-Armstrong asks, how is such an account to be assessed? Given that we can't actually see directly which counterfactuals people regard as relevant, how can we know whether the account is true or false?

This is exactly the right question to be asking, and I am sure that future research will offer us certain new techniques for answering it. At present, though, we have two major methods at our disposal.

First, the account predicts a particular pattern of intuitions across a broad range of different concepts. At the very heart of the approach is the idea that we should, as far as possible, avoid introducing ad hoc hypotheses just to explain the impact of moral judgment on one or another particular concept. Instead, we start out with perfectly general principles about the impact of moral judgment on counterfactual thinking. Then we introduce independently testable claims about the role of counterfactual thinking in the application of certain individual concepts. Together, these two types of claims generate specific testable predictions.

The thing to notice about this strategy is that it allows us to make predictions about the impact of moral considerations on the application of numerous concepts that have not yet been empirically investigated. Thus, to take one example, Jonathan Phillips (personal communication) points out that counterfactual reasoning seems to play a role in people's ordinary notion of *choosing*. (An agent cannot be said to have "chosen" one specific option unless other options were also available.) Hence, we should immediately predict an impact of moral judgment on people's intuitions about whether or not an agent can truly be said to have "chosen" a particular option. Or, to take a different case, it seems that counterfactual reasoning plays a role in people's intuitions about whether a given trait is *innate*. Accordingly, one might predict an impact of moral judgments on intuitions about innateness, and Richard Samuels and I are testing that prediction in a series of studies under development now. In essence, then, the first answer to **Sinnott-Armstrong**'s question is that we can test the theory by using it to generate new predictions about the application of various concepts and checking to see whether those predictions are borne out.

But there is also a second way in which the theory can be put to the test. We can use various methods to look more directly at people's judgments about the relevance of counterfactuals. For example, numerous studies have proceeded by presenting participants with questions of the form: "If only ___, this outcome would not have arisen." Participants can fill in the blank with whichever possibility they prefer, and researchers then infer that the possibilities chosen most often are regarded as most relevant. Studies using this methodology consistently show that moral judgments do have an impact on intuitions about counterfactual relevance (McCloy & Byrne 2000; N'gbala & Branscombe 1995).

In conclusion, then, our research can proceed by looking at the relationships among a complex constellation of different kinds of data. We start out with certain principles about the role of moral judgment in counterfactual thinking and certain hypotheses about the role of counterfactual thinking in the application of particular concepts. Then we check the theory against evidence regarding both counterfactual thinking and the application of concepts, testing to see whether all of these data conform to the theoretical predictions.[1] Presumably, they will not, and the theory will have to be revised in important respects. However, my hope is that we will at least be looking in roughly the right neighborhood and thereby moving toward a better understanding of these phenomena.

R2. The role of moral judgment

Suppose now that we focus, if only for the sake of argument, on the particular account advanced in the target article. The most important and controversial aspect of this account is the role it assigns to moral judgment. Yet, it can prove surprisingly difficult even to say what that role is and why it should be controversial, much less to determine whether the account is right or wrong.

R2.1. Investigating the judgments themselves

To begin with, there is the question as to what we even mean by the phrase "moral judgment." When one first hears this phrase, one is naturally drawn to think of a specific sort of conscious event. One thinks, for example, of cases in which we focus in on a particular behavior, bring to bear a variety of different considerations, and then determine that an agent deserves moral blame or praise.

Now, conscious episodes like this certainly do take place, but it sounds a bit implausible to suppose that such episodes could somehow be exerting a pervasive impact on people's whole way of understanding the

https://doi.org/10.1017/S0140525X10001810

world. We quite often wonder whether, for example, a person has a particular intention, and it seems absurd to suppose that whenever we want to answer such a question, we have to start out by making a full-blown moral judgment.

There is, however, a way of interpreting the hypothesis on which this sense of absurdity dissolves. To get a feeling for the issue, consider the way we might proceed if someone suggested that people's whole way of understanding the world was shaped by *statistical* reasoning. Clearly, when one first turns to the topic of statistical reasoning, one imagines a particular sort of conscious episode. (One thinks, perhaps, of a person moving step-by-step through the computations involved in a formal analysis of variance.) But surely the claim is not that *this* sort of cognition is shaping our whole understanding of the world! Rather, the idea is that people go through a kind of immediate, automatic, non-conscious process and that this process is analogous in certain important respects to what people do when they are consciously conducting statistical analyses.

The claim under discussion here should be understood in more or less this same way. We are certainly not suggesting that people's conscious moral beliefs can somehow shape their whole understanding of the world (see Knobe 2007). Rather, the claim is that people make certain immediate, automatic, non-conscious moral appraisals and that these automatic appraisals then exert a surprising influence on the rest of their cognition.

With this basic framework in mind, we can now turn to a series of interesting suggestions from the commentators.

R2.1.1. Theory-of-mind and counterfactuals. The commentaries from **Guglielmo** and **Girotto, Surian, & Siegal** (**Girotto et al.**) point to two important characteristics of people's moral judgments:

1. Guglielmo notes that conscious moral judgments are based in part on reasoning about the agent's mental states.
2. Girotto et al. note that conscious moral judgments are based in part on counterfactual reasoning.

These two points appear to spell trouble for the theory presented in the target article. After all, the claim was that people make a moral judgment which then influences their reasoning about mental states and counterfactuals. But if people have to think about mental states and counterfactuals before they can even make this moral judgment, how could the process ever get off the ground?

My answer is that the initial judgment that influences people's subsequent reasoning is deeply different from the conscious judgment that this reasoning can ultimately inform. People's conscious moral judgments can take into account information about numerous different considerations, including mental states, counterfactuals, and a great deal else besides. But their initial, purely non-conscious judgments do not work like that. These initial judgments are instead the product of an extremely rapid and far less complex process.

To see the basic idea here, imagine what might go through your mind if you were actually in the room as the vignette about the professor and the pens unfolded. There you are, watching as the professor moves toward the desk and starts reaching for one of the pens. Ultimately, you might end up making a conscious moral judgment about this behavior. You might decide that the professor deserves blame for the problem that results, or that his act was morally wrong, or something of the kind. But before you can even begin any of this sophisticated reasoning, you might go through a more automatic, entirely non-conscious process of moral appraisal. As you see the professor reaching for the pens, you recognize that he is suppose to refrain from taking them, and you therefore conceptualize his action by comparing it to the behavior he was supposed to perform, namely, refraining from taking pens. The key claims now are that (a) your tendency to focus on this specific comparison involves a kind of very simple moral cognition and (b) this simple form of moral cognition does not itself depend on your subsequent reasoning about mental states or counterfactuals.

R2.1.2. Origins of moral judgment. A question now arises about how exactly people make these rapid and automatic moral judgments. Here a number of commentators have provided helpful suggestions.

Kang & Glassman propose that moral judgments are shaped by the aim of acquiring cultural capital. People seek to signal their membership in particular communities and end up arriving at moral judgments accordingly. (Just as one might wear skinny jeans to signal one's membership in the community of Brooklyn hipsters, one might condemn abortion to signal one's membership in the community of Southern evangelicals.)

Terroni & Fraguas suggest that people's moral judgments can be impacted by their emotional states. They then hypothesize that people might make substantially different moral judgments when their emotional states were altered by clinical depression. So a person might arrive at different judgments about the very same case depending on whether that person happened to be depressed or not.

Carpendale, Hammond, & Lewis (**Carpendale et al.**) argue that people's capacity for moral judgment develops in the context of social interaction. Children learn to treat others as human beings (as opposed to mere physical objects), and they thereby acquire an understanding of moral norms.

Each of these hypotheses seems plausible and promising, but it would be especially exciting if we could use these approaches to drive a wedge between people's conscious moral judgments and their more automatic moral appraisals. Thus, suppose that an individual is trying to gain cultural capital by signaling membership in the community of liberal intellectuals. She might thereby end up arriving at the obvious sorts of conscious moral judgments: opposition to sexism and homophobia, support for disadvantaged groups, and so forth. But would her non-conscious appraisals go in this same way? Perhaps not. It might be that her conscious moral judgments would be shaped by the aim of gaining cultural capital, whereas her intuitions about intentional action, causation, and the like would continue to reveal a very different system of values at work (see, e.g., Inbar et al. 2009). Or consider the case of depression. Even when a person is clinically depressed, she may be able to exert enough cognitive control to continue making exactly the same sorts of conscious judgments that she would have otherwise. But perhaps her depression would nonetheless impact her non-conscious appraisals, and we might be able to pick up this impact just by asking questions about intention or causation.

R2.2. Impact of non-moral considerations

The commentaries from **Girotto et al.** and **Guglielmo** point out that people's intuitions about intentional action can be influenced, not only by moral considerations, but also by information about the agent's mental states. Thus, people are reluctant to say that an agent brought about an outcome intentionally when the agent shows regret (Guglielmo & Malle, in press; Phelan & Sarkissian 2008; Sverdlik 2004) or when the agent falsely believed that she would not be bringing the outcome about (Pellizzoni et al. 2010).

These are good points, and any correct theory of intentional action ascription will have to accommodate them. The theory presented in the target article does so by suggesting that moral considerations are used to set a kind of threshold, while information about the agent's mental states is used to determine whether the agent falls above or below that threshold. Hence, the position of the agent relative to the threshold ends up depending on a complex combination of moral considerations and mental state information.

R2.3. Moral concepts

What we have here, then, is a concept whose application can be influenced both by moral considerations and by mental state information. How should such a concept be understood? **Gintis** suggests that the best interpretation might be that people are simply using the concept of intentional action as a *moral concept*. The whole effect would then be rather unsurprising and unimportant. All it would show is that moral considerations can impact the application of moral concepts.

At least initially, this does seem like an appealing strategy. One starts out with a distinction between "moral" concepts and "non-moral" concepts, such that any concept whose application is impacted by moral considerations is supposed to fall in the former category. If one then finds an impact of moral considerations on a concept that had previously been classified as non-moral, one should not conclude that the whole framework is thereby called into question. All one needs to do is just reclassify that one concept.

Still, it does seem that there is a certain point at which this sort of strategy begins to look unworkable. If we find an impact of moral considerations on just one concept, we can always proceed by reclassifying it. But that is not the situation in which we actually find ourselves. These effects are arising not only for the concept of intentional action, but also for the concepts of causation and knowledge, even for the concept of advocating. At some point, I think, one has to conclude that it is becoming unhelpful to divide off a special sphere of "moral concepts" and claim that the impact of moral considerations arises only for them.

R2.4. Morality and normality

Kreps & Monin and **Mandelbaum & Ripley** take things even further in this direction. They suggest that the representation that is influencing people's intuitions in these cases is not actually specific to morality in particular. Rather, it is a representation of something like "normality" or "expectation." Such a representation would then unite moral considerations with considerations of a more purely statistical variety.

Continuing with this general approach, **Ulatowski & Johnson** propose that one can impact the relevant representation simply by creating stimulus materials that present a given outcome as a "default." Even if this outcome is not described as in any way morally good or right, the claim is that it will nonetheless be seen as having a particular sort of status that will prove relevant in people's subsequent cognition.

I think that the commentators are exactly right on this score, and I look forward to further research expanding on this theme. My only disagreement, if it can be considered a disagreement at all, is on the level of rhetoric. The commentators see themselves as deflating the claims made in the target article, showing that moral considerations are actually less central than I had originally suggested. By contrast, I would describe them as *radicalizing* the target article's original thesis. What they are showing is that it is not even possible to isolate a particular point in the process where the moral judgments come in. Instead, moral and statistical considerations appear to be welded seamlessly together from the very beginning.

R2.5. Morality and language

However, a number of commentators actually suggested moving in the opposite direction. They proposed theories according to which moral considerations are confined to a single, highly delimited role, while the remainder of the process has nothing to do with morality and proceeds more or less like a scientific investigation.

In particular, **Egré** and **Cova, Dupoux, & Jacob** (**Cova et al.**) suggest that the role of moral considerations might be confined entirely to *language*. The basic idea here is a simple and powerful one. Suppose that people's actual capacity for theory-of-mind works by classifying attitudes along a continuous scale. Still, it might be that our language cannot describe attitudes in these purely continuous terms. If we want to capture an agent's attitude in language, we need to impose some kind of threshold and then say that a particular term or phrase applies whenever the attitude goes beyond this threshold. So perhaps it is there that morality enters into the picture. In other words, it might be that the underlying scale is entirely non-moral, but that morality plays a role in the process we use to determine the position of the threshold for particular linguistic expressions.

One way of spelling out this sort of account would be to represent the underlying scale using numbers. We could say that the number 0 stands for absolute indifference, the positive numbers stand for *pro*-attitudes, and the negative numbers for *con*-attitudes. A particular agent's attitude could be represented using the diagram shown in Figure R1:

Yet, although people would have some representation of the agent's attitude along this scale, the actual expressions of natural language would not correspond to points on the scale in any absolute sense. So there would not be any expressions in English that could describe an agent as having an attitude of, say, "+2 or higher." Instead, all of the expressions of natural language would stand in a

https://doi.org/10.1017/S0140525X10001810

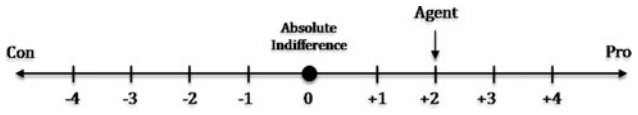

Figure R1. Representation of an agent's attitude on an absolute scale.

more complex relationship to the scale. They would characterize the agent's attitude *relative to* a (partially moral) default. Thus, if it turned out in the case at hand that the default was to an attitude of -1, the expressions of our language would describe the agent's attitude only relative to this default position, characterizing it as "3 points past the default."

There is, however, another possible way in which this system could work. It could be that human beings do not make use of any purely absolute representations at any stage of processing. Instead, the attitude would be represented from the very beginning in terms of its position relative to the default. We would start by labeling the 0 point as default and then represent the agent's attitude like this (Fig. R2):

On this latter view, the comparison with the default is already available in people's underlying, nonlinguistic representation of the attitude. The expressions of natural language can then correspond in a straightforward way to these nonlinguistic representations.

The primary difference between these two hypotheses is that the first posits an entirely non-moral representation, which is then obscured in certain ways by complex linguistic rules, whereas the second does not posit any purely non-moral representation at any level. The key to adjudicating between these hypotheses, then, is to come up with a richer account of what the non-moral representation is supposed to be doing. Given that it is not necessary as an explanation for the way people use expressions in natural language, what exactly *is* it used for? If we had a better account of what the non-moral representation was supposed to be doing, we would be better able to decide whether it is actually there.

R2.6. Characterizing the effect

The target article claims that moral considerations play a surprisingly important role in people's cognition. In trying to characterize this role, I adopted a number of different formulations. Sometimes I said that moral considerations figure in people's *competence*, sometimes that moral considerations suffuse the process *through and through*. The commentators suggest that both of those formulations are misleading and unhelpful.

Alexander, Mallon, & Weinberg (**Alexander et al.**) point out that no clear criteria are ever given for picking out a "competence" and distinguishing it from the various other factors that impact people's intuitions. They therefore suggest that we dispense with this distinction between competence and other factors and simply focus on exploring the various different processes that impact people's intuitions.

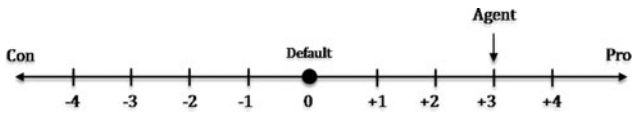

Figure R2. Representation of an agent's attitude relative to a default.

Stich & Wysocki note that there is a perfectly clear sense in which my own account does not have moral considerations influencing the process "through and through." On the contrary, the account says that moral considerations play a role in one specific part of the process but do not exert any influence on certain other parts.

These are both excellent points, and I agree that the formulations adopted in the target article may indeed be unhelpful in certain respects. So instead of defending what I wrote there, let me simply accept these criticisms and try now to formulate the point more accurately.

My aim in the target article was to argue against a particular vision. This vision distinguishes two aspects of the processes that generate people's intuitions:

1. A kind of "underlying" or "fundamental" capacity
2. Various additional factors that in some way "bias" or "distort" people's intuitions

The claim, then, is that the fundamental capacity is entirely non-moral and that the impact of moral considerations only arises because of the presence of these distorting factors.

Now, the distinction between these two aspects might be spelled out in various different ways, and different researchers would presumably adopt quite different accounts of the distinction. What unites all of these various accounts, however, is the claim that we can carve off a distinct capacity that is entirely non-moral and that is capable, all by itself, of generating an answer to the issue in question. Hence, faced with a person's intuition about intentional action, we might say: "This person's fundamental capacity for theory-of-mind would normally have classified this behavior as unintentional. However, her moral judgments got in the way and led her to regard it as intentional."

My aim was to show that this sort of strategy cannot be made to work. On the view I develop, there simply is no distinct capacity that is entirely non-moral and that is capable, all by itself, of determining whether a behavior is intentional or unintentional. Thus, on the model provided in the target article, there would be no sense in asking a question like: "Suppose we got rid of all the moral considerations and just allowed people's fundamental capacity for theory-of-mind to proceed undisturbed. Which conclusion would they then draw about whether this behavior was intentional?" The trouble here is that the model does not involve any kind of distinct non-moral capacity which could answer the question in the absence of all moral considerations.

Note that this argument does not require me to say anything positive about the distinction between competence and performance. Nor does it require me to claim that there is no stage anywhere in the process that is untouched by moral considerations. All it requires is a kind of negative claim. Specifically: that it not be possible to isolate a distinct capacity that has a particular sort of non-moral character.

R3. Ordinary cognition and science

In thinking about people's ordinary ways of making sense of the world, it sometimes proves helpful to draw analogies with more systematic and explicit systems of thought. So one might say that people's ordinary understanding is similar in certain respects to Aristotelian metaphysics, or to legal theory, or to certain religious doctrines. These analogies can then help to illuminate aspects of this ordinary understanding that might otherwise have remained obscure.

One particularly common analogy here has been between people's ordinary understanding and systematic *science*. This analogy calls up a specific picture of what the human mind is like. A scientific researcher might have two different kinds of beliefs in a particular domain – a system of scientific beliefs and then, quite separately, a system of moral beliefs. Such a researcher might then find that her collaborators strongly disagree with her moral beliefs but that they are nonetheless in complete agreement with her scientific beliefs.

In the target article, I argued that this analogy was misleading. People's ordinary cognition does not appear to involve a clear distinction between purely "scientific" beliefs and moral beliefs. It might be helpful, therefore, to reject the analogy with science and to look instead at analogies between ordinary cognition and forms of systematic thought in which moral and non-moral considerations are more explicitly mixed.

R3.1. The relevance of moral considerations

Spurrett & Martin argue that there is little to be gained by discussing the respects in which ordinary cognition might or might not resemble science. Instead, they suggest that we simply focus directly on the ways in which people apply specific considerations to address particular questions. Adopting this latter approach, they claim that the effects described in the target article are best characterized as "fallacies of relevance." That is, these effects are best understood as cases in which people apply moral considerations to questions in which only non-moral considerations would be relevant.

Spurrett & Martin may turn out in the end to be right on this score, but it is important to emphasize that the claim they are making is precisely the claim that is up for debate here. The central thesis of the target article was that people's ordinary cognition is radically different from scientific inquiry and that, in particular, ordinary questions like "Who caused the problem?" are not best understood on the model of scientific questions about causal relations. So, on the view defended in the target article, moral considerations actually *are* relevant to the ordinary questions people ask about whether one thing caused another, and there is no fallacy involved in applying such considerations to questions like these.

R3.2. Science and development

Kushnir & Chernyak suggest that the analogy to science might apply not so much to the beliefs people have at any given time but rather to the *development* of these beliefs in the first place. Hence, the beliefs people hold as adults might be radically different in various respects from the beliefs held by trained scientists, but the process people go through as children to acquire those beliefs might turn out to show many of the stages characteristic of scientific inquiry: looking for evidence, checking its fit to existing views, modifying these views when they do not fit the evidence, and so forth.

Kushnir & Chernyak's reference to the developmental literature here is a very helpful one, and future research could examine these developmental issues more directly. But it seems important at the outset to emphasize the very distinctive claim one makes when one says that ordinary human cognition resembles *science*. Such a claim presumably is not merely saying that ordinary human cognition involves taking in evidence and using it to assess prior views (a claim which is obviously true and needs no further defense). Instead, the claim seems to be an interesting and controversial one, which says something in particular about the precise way in which human beings use evidence to update their beliefs.

To see why, consider the way we might apply a similar approach in another domain. Suppose that someone says, "Human visual cognition uses Fourier transforms." The claim here is presumably not just that human visual cognition uses some kind of computation. Rather, what is being claimed is that visual cognition makes use of one specific kind of computation – a kind of computation that was first formalized by modern mathematicians and is now known as a Fourier transform. This is an interesting hypothesis, which can be put to the test in further experimental studies.

Now suppose that someone says: "Human cognitive development uses the methods of science." In just the same way, this claim cannot simply mean that cognitive development involves taking in evidence and using it to adjust our beliefs. (After all, that basic approach long predates the development of systematic science and can be found in an enormous variety of quite different modes of thought.) Rather, the claim has to be understood as saying that cognitive development makes use of the sorts of methods, first made explicit in the "scientific revolution" of the sixteenth and seventeenth centuries, that are now regarded as the distinctive methods of science. This is certainly an interesting hypothesis, which we can set about testing in experimental studies.

The thesis of the target article, however, was that existing experiments do not suggest that this hypothesis is correct. If we look to the distinctive characteristics of science – the characteristics that distinguish science from other systematic modes of thought – we find that people's ordinary non-conscious cognition *does not* tend to show these characteristics. For that reason, it might be helpful to understand ordinary cognition, not by looking to an analogy with contemporary science, but by looking to an analogy with the earlier modes of thought that the scientific revolution displaced.

R3.3. The function of theory-of-mind

Yet, even if the methods people use in ordinary theory-of-mind turn out to be radically different from the one we find at work in science, the *function* of theory-of-mind might be exactly the same as the function of scientific psychology. Thus, it might be that people's ordinary theory-of-mind makes use of moral considerations, but that

https://doi.org/10.1017/S0140525X10001810

there is some sense in which its aim is simply to generate accurate predictions and explanations of behavior.

Exploring this possibility, **Bartsch & Young** suggest that the impact of moral considerations might be understood in terms of information about frequency or probability. Suppose people generally assume that morally bad behaviors are infrequent or improbable. The judgment that a behavior was morally bad would then impact their statistical understanding, which could in turn influence their intuitions about intention, causation, and the like.

A number of other commentators take up related themes. **Baldo & Barberousse** propose that affective reactions can themselves serve as information and that this information may influence people's intuitions. And **Lombrozo & Uttich** point out that, even if moral considerations are entering into people's judgments at the algorithmic level, the best description at the computational level might still be in terms of an attempt to predict and explain behavior.

Now, it certainly does seem to be the case that people can sometimes use moral judgments to arrive at statistical truths, and these proposals therefore merit closer analysis. We should distinguish, however, between two possible ways in which the proposals might be interpreted.

One possible claim would be about the actual cognitive process people go through on-line. It might be claimed, for example, that people make a moral judgment, then use this judgment to make an inference about the frequency of the relevant behaviors, which in turn influences their intuitions about causation. If the proposal is understood in this way, I think that it is probably false. The problem is that when researchers independently vary information about frequency and moral status, they continue to find that moral status is playing an independent role (Roxborough & Cumby 2009).

But perhaps there is another, very different way of understanding the proposal. One might say that facts about frequencies are playing a role, not at the level of people's on-line cognition, but rather at the level of an "ultimate" or "evolutionary" explanation. Thus, suppose that theory-of-mind evolved as a mechanism for predicting and explaining behavior. Then, if violations of moral norms were generally infrequent, knowing that a behavior violated a norm might be a good cue for making certain statistical judgments about it, and our capacity for theory-of-mind might therefore have evolved to take moral considerations into account. In other words, the actual sequence of cognitive processes taking place in people's minds might involve all sorts of irreducibly moral appraisals, but the best evolutionary explanation of this process might be that it generally serves to enable accurate prediction. (For an especially clear defense of this approach, see the commentary by **Lombrozo & Uttich**.)

What we have here is a quite interesting hypothesis, but it is hard to know exactly how one might put it to the test empirically. In essence, we are dealing with a conflict between two very different visions. One vision focuses specifically on the nature of people's capacity for theory-of-mind. It says that this capacity has a particular "purpose" or "function" – for example, to accurately predict and explain behavior – and the patterns of intuition under discussion here can be explained in terms of their tendency to fulfill that function. By contrast, the vision I develop in the target article emphasizes certain general principles governing human cognition as a whole. The claim, then, is that the patterns we find in people's theory-of-mind judgments are not best understood as fulfilling any kind of purpose that is specific to theory-of-mind. Rather, these patterns simply reflect certain perfectly general principles about the impact of moral judgment on human cognition.

Clearly, the debate between these two visions is not the sort of thing that could be settled by a single critical experiment. Nonetheless, it does seem that further studies can offer us some illumination here. The key thing to notice is that the theory advanced in the target article predicts that the effects found in theory-of-mind should also be found in other domains that have nothing to do with theory-of-mind or even with prediction and explanation. So we can test the theory by looking to these other domains and checking to see whether similar effects are found there. An initial step in that direction can be found in the commentaries from **Egré** and from **Cova et al.**, both of which show an impact of moral judgment on the use of quantifiers like *many*. If we continue to find effects of that basic type, we will gradually acquire stronger and stronger reasons to conclude that the effects under discussion here are best explained in terms of very general facts about the structure of human cognition.

R3.4. The cognitive basis of science

Suppose, then, that people's ordinary way of making sense of the world really is deeply different from what one finds in systematic science. A question now arises about how the emergence of systematic science could even have been possible. Given that science is itself a human invention, how could the methods of science have ended up diverging so substantially from the methods characteristic of ordinary human cognition?

Levy offers a fascinating answer to this question. He suggests that the solution lies in the *social* character of science. In other words, the solution is not that each individual scientist can somehow enter a kind of special psychological state that allows her to transcend the limitations of ordinary human cognition and put all of her moral views to one side. Rather, the key point is that scientific inquiry is pursued by a whole community of different individuals, each of whom holds a slightly different set of moral views, and that this community as a whole is able to engage in a kind of inquiry that no single person could follow through on her own.

This suggestion strikes me as a deeply intriguing and promising one, and it would be wonderful to put it to the test in further experimental studies. Ideally, one would want to bring scientists into the lab and look systematically at the factors that influence their judgments. Assuming that scientists show many of the same effects found in lay people (e.g., Mercier & Sperber, forthcoming), there is good reason to expect that the presence of a broader community would have a substantial impact on their ability to call into question their own initial intuitions.

R4. Conclusion

Replies like this one are governed by some peculiar expectations. The author is supposed to fend off all the commentators' objections and show that his or her original article was actually completely correct all along. But, of course,

https://doi.org/10.1017/S0140525X10001810

I don't actually believe anything like that. A number of the hypotheses I defended in the past were subsequently refuted by other researchers, and I am sure that many of the hypotheses I have defended here will meet with a similar fate. Accordingly, it might be best to conclude, not by summarizing the views I hold right now, but rather by saying a few words about where things might move in the future.

When I first started investigating the impact of moral judgments on intuitions about intentional action, I assumed that most of people's cognition was entirely non-moral, and I therefore introduced a series of ad hoc maneuvers to explain the new experimental results. That strategy turned out to be completely misguided. As researchers began uncovering more and more cases in which morality influenced people's intuitions, it became ever more clear that we needed a theory that offered a more abstract characterization of the impact of morality on people's cognition as a whole.

I suspect that we will actually have to move even farther in that direction. As a number of the commentators noted, it might be a mistake to look for some special place where moral considerations enter the picture. Instead, we might need to develop a view on which the mind makes little distinction between moral and non-moral factors, so that the very same theory that explains the impact of moral considerations also explains our ability to make apparently "scientific" use of purely statistical or descriptive information.

NOTE

1. A quick note about the relevance of these data: The claim under discussion here is that judgments of counterfactual relevance play a role in intuitions about, e.g., causation. Hence, this claim yields the prediction that any factor that impacts judgments of counterfactual relevance should also impact intuitions about causation. In other words, if we uncover five different factors that influence judgments of counterfactual relevance, we should predict that all five of these factors influence causal intuitions, as well.

However, the claim does not also go the other way. We are not claiming that counterfactual thinking is the *only* thing that ever affects causal intuitions, so we are not claiming that every factor that influences causal intuitions must also influence counterfactual reasoning. On the contrary, as **Menzies** helpfully notes, a whole series of excellent studies have shown that people's causal intuitions can be influenced by factors that seem not to play a role in counterfactual thinking.

References

[The letters "a" and "r" before author's initials stand for target article and response references, respectively.]

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

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

Adams, F. & Steadman, A. (2007) Folk concepts, surveys, and intentional action. In: *Intentionality, deliberation, and autonomy: The action-theoretic basis of practical philosophy*, ed. C. Lumer, pp. 17–33. Ashgate. [aJK]

Alexander, J., Mallon, R. & Weinberg, J. (2010) Accentuate the negative. *Review of Philosophy and Psychology*. 1(2):297–314. [JA]

Alicke, M. (1992) Culpable causation. *Journal of Personality and Social Psychology* 63:368–78. [MA, rJK]

Alicke, M. (2000) Culpable control and the psychology of blame. *Psychological Bulletin* 126:556–74. [MA, aJK]

Alicke, M. (2008) Blaming badly. *Journal of Cognition and Culture* 8:179–86. [arJK]

Alicke, M., Rose, D. & Bloom, D. (2010) Causation, norm violation, and culpable control. Unpublished manuscript. [MA]

Alicke, M. & Zell, E. (2009) Social attractiveness and blame. *Journal of Applied Social Psychology* 39(9):2089–105. [MA]

Anscombe, G. E. M. (1958) *Intention*. Basil Blackwell. [TMS]

Asch, S. E. (1956) Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs* 70(9):1–70. [TAK]

Ashworth, A. (2006) *Principles of criminal law*, 5th edition. Oxford University Press. [FH]

Bechara, A., Tranel, D. & Damasio, H. (2000) Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain* 123:2189–202. [MVCB]

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

Benedetti F., Bernasconi, A., Blasi, V., Cadioli, M., Colombo, C., Falini, A., Lorenzi, C., Radaelli, D., Scotti, G. & Smeraldi, E. (2007) Neural and genetics correlates of antidepressant response to sleep deprivation: A functional magnetic resonance imaging study of moral valence decision in bipolar depression. *Archives of General Psychiatry* 64:179–87. [LT]

Biernat, M. (2005) *Standards and expectancies: Contrast and assimilation in judgments of self and others*. Psychology Press/Taylor and Francis. [TAK]

Bigler, R. S. & Liben, L. S. (2007) Developmental intergroup theory: Explaining and reducing children's social stereotyping and prejudice. *Current Directions in Psychological Science* 16:162–66. [TK]

Blair, R. J. R. (1996) Brief report: Morality in the autistic child. *Journal of Autism and Developmental Disorders* 26:571–79. [VG]

Blasi, A. (1980) Bridging moral cognition and moral action: A critical review. *Psychological Bulletin* 88:1–45. [MJK]

Bonawitz, E. B., Shafto, P., Gweon, H., Chang, I., Katz, S. & Schulz, L. (2009) The double-edged sword of pedagogy: Modeling the effect of pedagogical contexts on preschoolers' exploratory play. In: *Proceedings of the Thirty-first Meeting of the Cognitive Science Society*, ed. N.A. Taatgen & H. van Rijn, pp. 1575–80. (Online publication only, no publisher). [TK]

Bonnefon, J.-F. & Villejoubert, G. (2006) Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science* 17:747–51. [PE]

Bourdieu, P. (1986) The forms of capital. In: *Handbook of theory and research in the sociology of education*, ed. J. G. Richardson, pp. 241–58. Little, Brown. [MJK]

Bratman, M. (1987) *Intention, plans, and practical reason*. Harvard University Press. [FH]

Brogaard, B. (2010a) Adaptation, agency and intentional action. IRB-approved study at UM-SL, unpublish manuscript. [BB]

Brogaard, B. (2010b) The effects of personality assessment on judgments of intentional action. IRB-approved study at UM-SL, unpublished manuscript. [BB]

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

Bullock, M., Gelman, R. & Baillargeon, R. (1982) The development of causal reasoning. In: *The developmental psychology of time*, ed. W. J. Friedman, pp. 209–54. Academic Press. [TK]

Byrne, R. (2005) The rational imagination: How people create alternatives to reality. MIT Press. [aJK, PM]

Carey, S. & Spelke, E. (1996) Science and core knowledge. *Philosophy of Science* 63:515–33. [aJK]

Carpendale, J. I. M. & Lewis, C. (2004) Constructing an understanding of mind. *Behavioral and Brain Sciences* 27(1):79–150. [JIMC]

Carpenter, M., Akhtar, N. & Tomasello, M. (1998) Fourteen- to 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development* 21:315–30. [TK]

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

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

Cosmides, L., Barrett, H. C. & Tooby, J. (2010) Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Sciences USA* 107(Suppl. 2):9007–14. [NH]

Cova F. & Egré, P. (2010) Moral asymmetries and the semantics of "many." Unpublished manuscript, Institut Jean-Nicod. [PE]

Cushman, F. (2008) Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108:353–80. [SG]

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

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

Cushman, F. & Mele, A. (2008) Intentional action: Two-and-a-half folk concepts? In: *Experimental philosophy*, ed. J. Knobe & S. Nichols, pp. 171–88. Oxford University Press. [SG, aJK]

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. [aJK]

Darley, J. M. & Shultz, T. R. (1990) Moral rules: Their content and acquisition. *Annual Review of Psychology* 41:525–56. [SG]

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. [aJK]

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, pp. 307–38. Elsevier. [aJK]

Drevets, W. C., Price, J. L. & Furey, M. L. (2008) Brain structural and functional abnormalities in mood disorders: Implications for neurocircuitry models of depression. *Brain Structure and Function* 213(1–2):93–118. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18704495. [LT]

Driver, J. (2008a) Attributions of causation and moral responsibility. In: *Moral psychology, vol. 2: The cognitive science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong, pp. 423–40. MIT Press. [aJK]

Driver, J. (2008b) Kinds of norms and legal causation: Reply to Knobe and Fraser and Deigh. In: *Moral psychology, vol. 2: The cognitive science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong, pp. 459–62. MIT Press. [aJK]

Duff, R. A. (1990) *Intention, agency and criminal liability: Philosophy of action and the criminal law*. Basil Blackwell. [FH]

Duff, R. A. (1996) *Criminal attempts*. Oxford University Press. [FH]

Eagly, A. H., Wood, W. & Chaiken, S. (1978) Causal inference about communicators and their effect on opinion change. *Journal of Personality and Social Psychology* 36:424–35. [TAK]

Egré, P. (2010) Intentional action and the semantics of gradable expressions (On the Knobe Effect). Unpublished manuscript, Institut Jean-Nicod. [PE]

Eslinger, P. J. & Damasio, A. R. (1985) Severe disturbance of higher cognition after bilateral frontal lobe ablation: Patient EVR. *Neurology* 35:1731–41. [MVCB]

Fara, D. (2000) Shifting sands: An interest-relative theory of vagueness. *Philosophical Topics* 28(1):45–81. (Originally published under the name Delia Graff). [PE]

Feltz, A. & Cokely, E. T. (2007) An anomaly in intentional action ascriptions: More evidence of folk diversity. In: *Proceedings of the 29th Annual Cognitive Science Society*, ed., D. S. McNamara & J. G. Trafton, p. 1748. Cognitive Science Society. [aJK]

Fiske, A. P. (1992) Four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review* 99:689–723. [MVCB]

Gellatly, A. (1997) Why the young child has neither a theory of mind nor a theory of anything else. *Human Development* 40:32–50. [JIMC]

Girotto, V., Ferrante, D., Pighin, S. & Gonzalez, M. (2007) Post-decisional counterfactual thinking by actors and readers. *Psychological Science* 18:510–15. [VG]

Girotto, V., Legrenzi, P. & Rizzo, A. (1991) Counterfactual thinking: The role of events controllability. *Acta Psychologica* 78:111–33. [VG]

Glassman, M. (1996) Understanding Vygotsky's motive and goal: An exploration of the work of A. N. Leontiev. *Human Development* 39:309–27. [MJK]

Goldman, A. (1970) *A theory of human action*. Prentice-Hall. [aJK]

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

Gopnik, A. (1996) The scientist as child. *Philosophy of Science* 63:485–514. [MVCB]

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T. & Danks, D. (2004) A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review* 111:1–30. [aJK, TK]

Gopnik, A. & Meltzoff, A. (1997) *Words, thoughts and theories*. MIT Press. [aJK]

Gopnik, A. & Schulz, L. (2004) Mechanisms of theory formation in young children. *Trends in Cognitive Sciences* 8(8):371–77. [MVCB]

Gopnik, A., Sobel, D. M., Schulz, L. E. & Glymour, C. (2001) Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology* 37(5):620–29. [TK]

Gopnik, A. & Tenenbaum, J. B. (2007) Bayesian networks, Bayesian learning and cognitive development. *Developmental Science* 10(3):281–87. [MVCB]

Gopnik, A. & Wellman, H. M. (1992) Why the child's theory of mind really is a theory. *Mind and Language* 7:145–71. [KB, aJK]

Greene, J. D. (2008) The secret joke of Kant's soul. In: *Moral psychology, vol. 3: The neuroscience of morality: Emotion, disease, and development*, ed. W. Sinnott-Armstrong, pp. 35–79. MIT Press. [SG]

Greene, J. & Haidt, J. (2002) How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6(12):517–23. [MVCB]

Grice, H. P. (1975a) Logic and conversation. In: *Syntax and semantics, vol. 3: Speech acts*, ed. P. Cole & J. L. Morgan. Academic Press. [JIMC]

Grice, H. P. (1975b) Logic and conversation. In: *The logic of grammar*, ed. D. Davidson & G. Harman, pp. 64–75. Dickenson. [HG]

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

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

Guglielmo, S. & Malle, B. F. (in press) Can unintended side-effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*. [SG, arJK]

Guglielmo, S., Monroe, A. E. & Malle, B. F. (2009) At the heart of morality lies folk psychology. *Inquiry* 52:449–66. [SG]

Haidt, J. (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108:814–34. [MVCB, KB, SG, aJK]

Haidt, J. (2007) The new synthesis in moral psychology. *Science* 316:998–1002. [MVCB]

Hamlin, J. K., Wynn, K. & Bloom, P. (2007) Social evaluation by preverbal infants. *Nature* 450:557–59. [KB, TK]

Hart, H. L. A. & Honoré, A. M. (1985) *Causation in the law*, 2nd edition. Oxford University Press. [PM]

Heyman, G. & Gelman, S. (2000) Beliefs about the origins of human psychological traits. *Developmental Psychology* 36(5):663–78. [TK]

Hindriks, F. (2008) Intentional action and the praise-blame asymmetry. *Philosophical Quarterly* 58:630–41. [FH, arJK]

Hitchcock, C. & Knobe, J. (2009) Cause and norm. *Journal of Philosophy* 106(11):587–612. [KB, SG, arJK, PM]

Hoffman, M. L. (2000) *Empathy and moral development.* Cambridge University Press. [KB]

Holiday, A. (1988) *Moral powers: Normative necessity in language and history*. Routledge. [JIMC]

Hume, D. (1739/2000) *A treatise of human nature*. Oxford University Press. [MVCB]

Inbar, Y., Pizarro, D. A., Knobe, J. & Bloom, P. (2009) Disgust sensitivity predicts intuitive disapproval of gays. *Emotion* 9(3):435–39. [arJK, EM]

Kahneman, D. & Miller, D. (1986) Norm theory: Comparing reality to its alternatives. *Psychological Review* 93:136–53. [VG, aJK]

Kalish, C. (2002) Children's predictions of consistency in people's actions. *Cognition* 84(3):237–65. [TK]

Kang, M. & Glassman, M. (2010) Moral action as social capital, moral thought as cultural capital. *Journal of Moral Education* 39:21–36. [MJK]

Kelley, H. H. (1967) Attribution theory in social psychology. In: *Nebraska Symposium on Motivation*, ed. D. Levine, pp. 192–238. University of Nebraska Press. [aJK]

Kelley, H. H. (1971) *Attribution in social interaction*. General Learning Press. [TAK]

Kennedy, C. (2007) Vagueness and grammar: The semantics of absolute and relative gradable adjectives. *Linguistics and Philosophy* 30:1–45. [PE]

Kitcher, P. (1993) *The advancement of science: Science without legend, objectivity without illusions.* Oxford University Press. [NL]

Knobe, J. (2003a) Intentional action and side effects in ordinary language. *Analysis* 63:190–93. [aJK, BN, JU]

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

Knobe, J. (2004a) Folk psychology and folk morality: Response to critics. *Journal of Theoretical and Philosophical Psychology* 24(2):270–79. [aJK]

Knobe, J. (2004b) Intention, intentional action and moral considerations. *Analysis* 64:181–87. [aJK]

Knobe, J. (2006) The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies* 130:203–31. [BB, aJK]

Knobe, J. (2007) Reason explanation in folk psychology. *Midwest Studies in Philosophy* 31:90–107. [arJK, BN]

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

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

Knobe, J. & Fraser, B. (2008) Causal judgment and moral judgment: Two experiments. In: *Moral psychology, vol. 2: The cognitive science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong, pp. 441–8. MIT Press. [aJK]

Knobe, J. & Mendlow, G. (2004) The good, the bad, and the blameworthy: Understanding the role of evaluative considerations in folk psychology. *Journal of Theoretical and Philosophical Psychology* 24:252–58. [EM]

Knobe, J. & Roedder, E. (2009) The ordinary concept of valuing. *Philosophical Issues* 19(1):131–47. [aJK]

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. [aJK]

Kunda, Z. (1990) The case for motivated reasoning. *Psychological Bulletin* 108(3):480–98. [aJK]

Kushnir, T., Wellman, H. M. & Gelman, S. A. (2007) The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition* 107(3):1084–92. [TK]

Kushnir, T., Wellman, H. M. & Gelman, S. A. (2009) A self-agency bias in children's causal inferences. *Developmental Psychology* 45(2):597–603. [TK]

Kushnir, T., Xu, F. & Wellman, H. M. (2010) Young children use statistical sampling to infer the preferences of others. *Psychological Science* 21:1134–40. [TK]

Lappin S. (2000) An intensional parametric semantics for vague quantifiers. *Linguistics and Philosophy* 23:599–620. [PE]

Leslie, A. M., Friedman, O. & German, T. P. (2004) Core mechanisms in "Theory of Mind." *Trends in Cognitive Sciences* 8:528–33. [VG]

Leslie, A. M. & Keeble, S. (1987) Do six-month-old infants perceive causality? *Cognition* 25:265–88. [TK]

Leslie, A. M., Knobe, J. & Cohen, A. (2006a) Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science* 17:421–07. [aJK]

Leslie, A. M., Mallon, R. & DiCorcia, J. A. (2006b) Transgressors, victims, and cry babies: Is basic moral judgment spared in autism? *Social Neuroscience* 1:270–83. [VG]

Lutz, D. & Keil, F. (2002) Early understanding of the division of cognitive labor. *Child Development* 73:1073–84. [TK]

Machery, E. (2008) The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language* 23(2):165–89. [SG, aJK, EM, BN]

Macrae, C. N. (1992) A tale of two curries: Counterfactual thinking and accident-related judgments. *Personality and Social Psychology Bulletin* 18:84–87. [VG]

Malinowski, B. (1922) *Argonauts of the Pacific*. Percy, Land, Humphries. [MJK]

Malle, B. & Nelson, S. (2003) Judging *mens rea*: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law* 21:563–80. [FH]

Mallon, R. (2007) Reviving Rawls inside and out. In: *Moral psychology, vol. 2: The cognitive science of morality: Intuition and diversity*, ed. W. Sinnott-Armstrong, pp. 145–55. MIT Press. [JA]

Mallon, R. (2008) Knobe vs. Machery: Testing the trade-off hypothesis. *Mind and Language* 23:247–55. [BN]

Mandel, D. R. (2003) Judgment dissociation theory: An analysis of differences in causation, counterfactual, and covariational reasoning. *Journal of Experimental Psychology: General* 137:419–34. [PM]

Mandel, D. R. & Lehman, D. R. (1996) Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology* 71:450–63. [PM]

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt. [TL]

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, February 27–28, 2009. [aJK]

McArthur, L. & Post, D. (1977) Figural emphasis and person perception. *Journal of Experimental Social Psychology* 13:520–35. [aJK]

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

McCloy, R. & Byrne, R. (2000) Counterfactual thinking about controllable events. *Memory and Cognition* 28:1071–78. [arJK]

Mead, G. H. (1934) *Mind, self and society*. University of Chicago Press. [JIMC]

Meltzoff, A. (1995) Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology* 31:838–50. [TK]

Menzies, P. (2007) Causation in context. In: *Causation, physics, and the constitution of reality: Russell's republic revisited*, ed. H. Price & R. Corry, pp. 191–223. Oxford University Press. [PM]

Menzies, P. (2009) Platitudes and counterexamples. In: *The Oxford handbook of causation*, ed. H. Beebee, C. Hitchcock & P. Menzies, pp. 341–67. Oxford University Press. [PM]

Mercier, H. & Sperber, D. (forthcoming) Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*. [rJK]

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

Moll, J., Zahn, R., Oliveira-Souza, R., Krueger, F. & Grafman, J. (2005) The neural basis of human moral cognition. *Nature Reviews Neuroscience* 6:799–809. [MVCB]

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

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* 31:139–62. [aJK]

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

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

Nadelhoffer, T. (2006b) Desire, foresight, intentions, and intentional actions: Probing folk intuitions. *Journal of Cognition and Culture* 6:133–57. [SG]

Nadelhoffer, T. (2006c) On trying to save the Simple View. *Mind and Language* 21:565–86. [aJK]

Nanay, B. (2010) Morality of modality? What does the attribution of intentionality depend on? *Canadian Journal of Philosophy* 40:28–40. [aJK, BN]

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* 31:139–62. [rJK]

Nichols, S. & Ulatowski, J. (2007) Intuitions and individual differences: The Knobe effect revisited. *Mind and Language* 22:346–65. [arJK, BN, JU]

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

Oakes, L. M. & Cohen, L. B. (1990) Infant perception of a causal event. *Cognitive Development* 5:193–207. [TK]

Pellizzoni, S., Girotto, V. & Surian, L. (2010) Beliefs and moral valence affect intentionality attributions: The case of side effects. *Review of Philosophy and Psychology* 1:201–209. [SG, VG, rJK]

Pellizzoni, S., Siegal, M. & Surian, L. (2009) Foreknowledge, caring, and the side-effect effect in young children. *Developmental Psychology* 45:289–95. [VG]

Pettit, D. & Knobe, J. (2009) The pervasive impact of moral judgment. *Mind and Language* 24:586–604. [FC, PE, aJK, TAK]

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(2):291–98. [SG, arJK, EM]

Phillips, J. & Knobe, J. (2009) Moral judgments and intuitions about freedom. *Psychological Inquiry* 20:30–36. [aJK]

Piaget, J. (1932/1965) *The moral judgment of the child*. The Free Press. (Original work published in 1932). [JIMC]

Pighin, S., Bonnefon, J.-F. & Savadori, L. (2009) Overcoming number numbness in prenatal risk communication. Unpublished manuscript, University of Toulouse and Department of Cognitive Science and Education, University of Trento. [PE]

Pizarro, D., Uhlmann, E. & Salovey, P. (2003) Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science* 14:267–72. [SG]

Portes, A. (1998) Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology* 24:1–24. [MJK]

Premack, D. & Woodruff, G. (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1:515–26. [aJK]

Prentice, D. A. & Miller, D. T. (1996) Pluralistic ignorance and the perpetuation of social norms by unwitting actors. *Advances in Experimental Social Psychology* 28:161–210. [TAK]

Prinz, J. (2006) The emotional basis of moral judgment. *Philosophical Explorations* 9:29–43. [MVCB]

Pugliucci, M. (2010) *Nonsense on stilts: How to tell science from bunk*. University of Chicago Press. [NH]

Putnam, R. D. (2001) *Bowling alone: The collapse and revival of American community*. Simon & Schuster. [MJK]

Reeder, G. D. & Brewer, M. B. (1979) A schematic model of dispositional attribution in interpersonal perception. *Psychological Review* 86:61–79. [SG]

Repacholi, B. & Gopnik, A. (1997) Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology* 33(1):12–21. [TK]

Rhodes, M. & Gelman, S. A. (2008) Categories influence predictions about individual consistency. *Child Development* 79:1271–88. [TK]

Rhodes, M., Gelman, S. A. & Brickman, D. (in press) Children's attention to sample composition in learning, teaching, and discovery. *Developmental Science*. [TK]

Roese, N. (1997) Counterfactual thinking. *Psychological Bulletin* 121:133–48. [aJK]

Ross, L. & Ward, A. (1996) Naive realism in everyday life: Implications for social conflict and misunderstanding. In: *Values and knowledge*, ed. E. Reed, E. Turiel, & T. Brown, pp. 103–35. Erlbaum. [TAK]

Roxborough, C. & Cumby, J. (2009) Folk psychological concepts: Causation. *Philosophical Psychology* 22:205–13. [KB, arJK]

Sapir E. (1944) Grading: A study in semantics. *Philosophy of Science* 11(2):93–116. [PE]

Saxe, R., Tzelnic, T. & Carey, S. (2007) Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. *Developmental Psychology* 43(1):149–58. [TK]

Scanlon, T. M. (2008) *Moral dimensions: Permissibility, meaning, blame*. Harvard University Press. [TMS]

Schulz, L. E. & Bonawitz, E. B. (2007) Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology* 43(4):1045–50. [TK]

Schulz, L. E., Kushnir, T. & Gopnik, A. (2007) Learning from doing: Interventions and causal inference. In: *Causal learning: Psychology, philosophy and computation*, ed. A. Gopnik & L. E. Schulz, pp. 67–86. Oxford University Press. [TK]

Setiya, K. (2003) Explaining action. *Philosophical Review* 112:339–93. [FH]

Shaver, K. G. (1985) *The attribution of blame: Causality, responsibility, and blameworthiness*. Springer. [SG]

Sheline, Y. I., Barch, D. M., Price, J. L., Rundle, M. M., Vaishnavi, S. N., Snyder, A. Z., Mintun, M. A., Wang, S., Coalson, R. S. & Raichle, M. E. (2009) The default mode network and self-referential processes in depression. *Proceedings of the National Academy of Sciences USA* 106(6):1942–47. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19171889. [LT]

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

Shimizu, Y. A. & Johnson, S. C. (2004) Infants' attribution of a goal to a morphologically unfamiliar agent. *Developmental Science* 7(4):425–30. [TK]

Shultz, T. R. (1982) Rules of causal attribution. *Monographs of the Society for Research in Child Development* 47, Serial No. 194. [TK]

Siever, E., Gopnik, A. & Goodman, N. (under review) Did she jump because she was brave or because the trampoline was safe? Causal inference and the development of social cognition. [TK]

Sinnott-Armstrong, W. (2008) A contrastivist manifesto. *Social Epistemology* 22(3):257–70. [WS-A]

Skowronski, J. J. & Carlston, D. E. (1989) Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin* 105:131–42. [SG]

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

Smedslund, J. (1963) The concept of correlation in adults. *Scandinavian Journal of Psychology* 4:165–73. [aJK]

Sobel, D. & Kirkham, N. (2006) Bayes nets and babies: Infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science* 10(3):298–306. [TK]

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

Spelke, E., Phillips, A. & Woodward, A. (1995) Infants' knowledge of object motion and human action. In: *Causal cognition: A multidisciplinary debate*, pp. 44–78. Clarendon Press/Oxford University Press. [TK]

Spelke, E. S. & Kinzler, K. D. (2007) Core knowledge. *Developmental Science* 10:89–96. [KB]

Stocker, M. (1973) Act and agent evaluations. *Review of Metaphysics* 27:42–61. [FH]

Suddendorf, T. & Whiten, A. (2001) Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals. *Psychological Bulletin* 127:629–50. [BN]

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. [aJK]

Surian, L., Caldi, S. & Sperber, D. (2007) Attribution of beliefs by 13-month-old infants. *Psychological Science* 18:580–86. [VG]

Sverdlik, S. (2004) Intentionality and moral judgments in commonsense thought about action. *Journal of Theoretical and Philosophical Psychology* 24:224–36. [arJK]

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

Tetlock, P. E. (2002) Social-functionalist frameworks for judgment and choice: The intuitive politician, theologian, and prosecutor. *Psychological Review* 109:451–72. [aJK]

Thorndike, E. L. (1920) A constant error in psychological rating. *Journal of Applied Psychology* 4:25–29. [DS]

Turiel, E. (1983) *The development of social knowledge: Morality and convention.* Cambridge University Press. [TK]

Turiel, E. (2006) The development of morality. In: *Handbook of child psychology, vol. 3: Social, emotional, and personality development*, 6th edition, ed. N. Eisenberg, pp. 789–857. Wiley. [KB]

Turnbull, W. (2003) *Language in action: Psychological models of conversation.* Psychology Press. [JIMC]

Turner, J. (2004) Folk intuitions, asymmetry, and intentional side effects. *Journal of Theoretical and Philosophical Psychology* 24:214–19. [aJK]

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

Ulatowski, J. & Johnson, J. (2010) Folk intuitions and Quinn's doctrine of doing and allowing. Unpublished manuscript, University of Nevada, Las Vegas. [JU]

Uttich, K. & Lombrozo, T. (2010) Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition* 116:87–100. [SG, VG, TL]

Velleman, J. D. (1989) *Practical reflection*. Princeton University Press. [FH]

Weber, E. & Hilton, D. J. (1990) Contextual effects in the interpretation of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance* 16(4):781–89. [PE]

Wellman, H. M. (1990) *The child's theory of mind*. MIT Press. [TK]

Winch, P. (1972) *Ethics and action*. Routledge and Kegan Paul. [JIMC]

Woodward, A. (1998) Infants selectively encode the goal object of an actor's reach. *Cognition* 69(1):1–34. [TK]

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

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

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. [arJK]

Young, L. & Saxe, R. (2009) Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia* 47:2065–72. [SG]

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