The Pervasive Impact of Moral Judgment[*]

Dean Pettit and Joshua Knobe

Forthcoming in *Mind & Language*

Abstract: A series of recent studies have shown that people's moral judgments can affect their intuitions as to whether or not a behavior was performed intentionally. Prior attempts to explain this effect can be divided into two broad families. Some researchers suggest that the effect is due to some peculiar feature of the concept of intentional action in particular, while others suggest that the effect is a reflection of a more general tendency whereby moral judgments exert a pervasive influence on folk psychology. The present paper argues in favor of the latter hypothesis by showing that the very same effect that has been observed for *intentionally* also arises for *deciding, in favor of, opposed to*, and *advocating*.

People ordinarily distinguish between behaviors that are performed 'intentionally' and those that are performed 'unintentionally.' At first glance, this distinction seems to be a perfectly familiar part of our ordinary approach to understanding the mind, right alongside the concepts of belief and desire. In other words, the concept of intentional action appears to be one aspect of *folk psychology*.

Yet recent experimental work has revealed a surprising fact about the way in which people ordinarily apply this concept. It seems that people's ordinary intuitions about intentional action can actually be affected by their *moral* judgments. In particular, there seem to be cases in which people's intuitions about whether a behavior was performed intentionally depend in some way on their moral appraisal of the behavior itself. What we have here, then, is a case in which people's moral judgments appear to be influencing their folk-psychological intuitions.

A question now arises as to whether this effect is telling us anything of general significance about the relationship between folk psychology and moral judgment. Is the effect just due to some quirk in the process by which people attribute intentional action, or is it a manifestation of some more general mechanism whereby moral judgments can

[*] We are grateful for comments on an earlier draft to Fiery Cushman, Edouard Machery, Rebecca Saxe and Liane Young and for theoretical inspiration to Nichols and Ulatowski (2007).

have an impact on folk psychology? Here, one finds a striking divergence of views – with researchers dividing off into two basic camps.

On one side are researchers who suggest that the effect can be understood entirely in terms of certain special features of the attribution of intentional action in particular (e.g., Machery 2008; Nichols and Ulatowski 2007). These researchers propose to explain the effect by positing a process that would apply only to attributions of intentional action and would not be expected to arise for any other aspect of folk psychology.

On the other side are researchers who think that the effect can be explained in terms of some very general fact about the relationship between folk psychology and moral judgment (e.g., Alicke forthcoming; Knobe 2006; Nadelhoffer 2006; Nado 2008). These researchers then proceed by constructing general theories about the ways in which moral judgments impact folk psychology. The guiding hope is that, if one can arrive at the correct general theory, the specific facts about intentional action will be seen to be just one aspect of a far broader pattern.

Our aim here is to provide experimental and theoretical support for this second view. On the theory we develop here, the surprising results obtained for intuitions about intentional action do not really have anything to do with the distinctive features of the concept of intentional action in particular. Rather there is a perfectly general process whereby moral judgments serve as input to folk psychology, and the effects observed for intentional action should be understood as just one manifestation of this broader phenomenon. If we are right about this, the impact of moral judgments is not merely a peculiarity of the concept of intentional action, but instead is a pervasive feature of the theory of mind.

Background

Consider a paradigmatic case of intentional action. The agent wants to bring about an outcome, she performs a behavior specifically for that purpose, and everything proceeds exactly as planned. In a case like this one, people's intuitions will be more or less independent of moral considerations. Regardless of whether the behavior is morally good or morally bad, almost everyone will say that the agent brought about the outcome intentionally.

Now consider a behavior that is paradigmatically unintentional. The agent has no interest in bringing about the outcome, she doesn't even know that her behavior might bring it about, and she only ends up acting as a result of some sort of muscle spasm. Here again, moral considerations will have little impact on people's intuitions. No matter what moral status the behavior has, almost everyone will say that the agent brings about the outcome unintentionally.

Things get interesting, however, when we consider intermediate cases – i.e., cases that fall somewhere between the paradigmatically intentional and the paradigmatically unintentional. Thus, suppose that the agent knows that she will be bringing about a particular outcome through her behavior but that she does not care about this outcome in any way. (She has chosen to perform the behavior for some other reason entirely.) In such a case, we might say that the outcome is a 'side-effect' of her behavior. Will people say that she brought about this side-effect intentionally? It turns out that their intuitions in cases of this type can actually be influenced by their judgments about whether the side-effect itself is morally good or morally bad.[1]

The usual way of demonstrating this influence of moral judgment on attributions of intentional action is to present experimental subjects with cases in which an agent brings about a side-effect that is either morally good or morally bad. Here, for example, is a case that we will call the *harm 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.

[1] Here we are simplifying for the sake of expository convenience. It is widely agreed that some sort of normative judgment is affecting people's intuitions, but no one now thinks that the relevant judgment is just a judgment about whether the side-effect is morally good or morally bad. At this point, the two major views are (a) that the effect is the result of a complex interplay among a number of different normative judgments (Machery forthcoming; Phelan and Sarkissian forthcoming; Wright and Bengson 2007) and (b) that none of our consciously-held normative judgments are playing a role and that the effect should be understood instead in terms of a fast, automatic, entirely non-conscious kind of normative appraisal (Alicke forthcoming; Knobe 2007; Nadelhoffer 2004). The details of this debate will not prove relevant to any of the questions we discuss here, and we therefore put the issue to one side.

After reading this vignette, subjects can be asked whether they agree or disagree with the statement: 'The chairman of the board intentionally harmed the environment.'

But now suppose we construct a case that is almost exactly the same as this first one, except that the side effect is actually morally good. We then arrive at what we will call the *help vignette*:

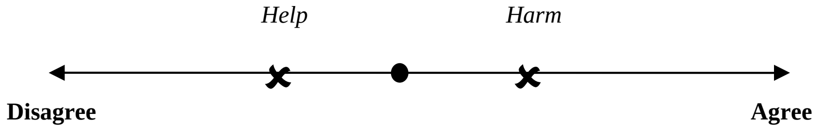
> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.'
>
> The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was helped.

After reading this second vignette, subjects can be asked whether they agree or disagree with the statement: 'The chairman of the board intentionally helped the environment.'

Experimental studies concerning intuitions about cases like these consistently show a striking asymmetry (Feltz and Cokely 2007; Knobe 2003; Mallon 2008; Nichols and Ulatowski 2007; Phelan and Sarkissian forthcoming). Subjects who receive the harm vignette typically say that the agent intentionally harmed the environment, whereas subjects who receive the help vignette typically say that the agent did not intentionally help the environment. Yet it seems that the agent's mental states do not differ between the two cases. The main difference lies instead in the moral status of the side-effect itself. Hence, most researchers have concluded that people's moral judgments are somehow influencing their intuitions as to whether or not an agent acts intentionally (Knobe 2006; Malle 2006; Nadelhoffer 2006; Nado 2008).

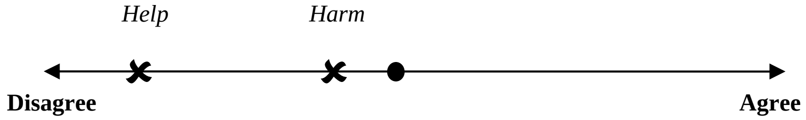
The key question now is whether this effect has something to do with the concept of intentional action in particular or whether it is simply one manifestation of a pervasive influence of moral judgment on folk psychology. In the experiment we have been discussing thus far, subjects were presented with the help and harm vignettes and asked in each case whether the agent acted *intentionally*, but what would have happened if they had instead been asked a question using some other folk-psychological concept? Suppose they had been asked whether the agent had a *desire* to help or harm the environment. Or suppose they had been asked whether the agent was *in favor* of helping or harming the environment. Would the effect then have disappeared? Or would we have found the very same asymmetry using those concepts as well?

Before we address these issues in earnest, a word is in order about the precise way in which we will be measuring levels of agreement and disagreement. In each of the experiments we report here, subjects are presented with a sentence and then asked to rate that sentence on a scale from 'disagree' to 'agree.' When subjects were given the sentences about helping and harming and asked whether the agent acted intentionally, their responses came out roughly as follows (Knobe 2005):



Note that the ratings for 'help' and 'harm' are actually on opposite sides of the midpoint, with the rating for 'help' on the side of disagreement and the rating for 'harm' on the side of agreement. It might be tempting, then, to suppose that the key result of the study is that subjects concluded on the whole that the agent acted intentionally in the harm vignette but unintentionally in the help vignette.

We think that this temptation should be resisted. For present purposes, the important thing is not whether people's responses fall on one or the other side of the midpoint but rather whether there is a *difference* between responses to the morally bad case and responses to the morally good case. After all, suppose that some other factor – a factor that had nothing at all to do with the issues under discussion in this paper – shifted both types of responses a little bit to the left. The two responses might then come out as follows:



In this latter case, responses in both conditions are to the left of the midpoint, and a certain approach to thinking about people's intuitions might leave us with the idea that the key result was simply that, on the whole, subjects were disagreeing in both conditions. But this sort of approach serves only to obscure what is most relevant here.

The important point is that moral judgments are influencing intuitions in this case in precisely the same way, and perhaps for the very same reasons, that moral judgments influence intuitions in the case described above.

Our approach will therefore be to focus not so much on the absolute levels of agreement in each case as on the differences between levels of agreement in morally good and morally bad cases. Using this basic approach, we can then examine the impact of moral judgment on people's application of an array of different concepts.

Evidence of Pervasiveness

When one pursues this research program, one quickly runs up against a surprising result. Not only does the impact of moral judgment extend beyond the concept of intentional action, moral judgments appear to be having some impact on just about every concept that involves holding or displaying a positive attitude toward an outcome. We will present data on six different concepts in this section, then turn to another two cases shortly thereafter.

1. 'Intention' and 'Intend'

One striking finding from recent work on the concept of intentional action is the surprising difference between people's use of the adverb 'intentionally' and their use of the verb 'intend' and the noun 'intention.' Perhaps the strongest evidence here comes from a study by McCann (2005) in which subjects were given the harm vignette and asked:

- Did the chairman intentionally harm the environment?
- Did the chairman intend to harm the environment?
- Was it the chairman's intention to harm the environment?

In that study, most subjects (64%) said that the agent acted 'intentionally,' but less than half (42%) said that he did 'intend' and relatively few (27%) said that he had an 'intention.'

At this point, one might conclude that morality does not have the same sort of effect on 'intend' and 'intention' that it does on 'intentionally.' (After all, the majority of subjects in the study are disagreeing with the claim that the agent 'intended' or had the 'intention.') But appearances here are misleading. While only a minority of subjects are applying these terms in the harm case, one can still see evidence of a moral asymmetry.

Thus, in one recent study (Knobe 2004), subjects were randomly assigned to receive either the help vignette or the harm vignette and then asked:

- Was it the chairman's intention to harm [help] the environment?

Although relatively few (29%) subjects said that the agent had an intention to harm, absolutely none (0%) said he had an intention to help. So people tended not to ascribe intention in either of these cases, but they were more likely to ascribe intention in the case where the behavior was morally bad.

Similar effects have been observed for the verb 'intend.' Cushman (2007) developed 21 different scenarios about agents who brought about side-effects. Each scenario was constructed with two versions – one in which the action is morally good, another in which the action is morally bad. In all 21 scenarios, subjects showed higher levels of agreement with the statement that the agent 'intended' to bring about the side-effect in the morally bad version than in the morally good version.

2. 'Desire'

Here one might suspect that the words 'intentionally,' 'intend' and 'intention' all express more or less the same concept and that the effect might disappear as soon as one turns to words that express other folk-psychological concepts. That, however, appears not to be the case. In fact, the effect also emerges when one looks at applications of 'desire.'

Tannenbaum, Ditto and Pizarro (2007) conducted a study in which subjects were presented with the help and harm vignettes and then asked:

- Did the chairman have a desire to help [harm] the environment?

Subjects marked their answers to this question on a scale from 1 to 7. The mean for the help vignette was 1.6; the mean for the harm vignette was 3.4. Here again, although subjects in both conditions leaned toward a negative answer to the question, subjects assigned significantly higher ratings in the morally bad case than in the morally good case.

3. 'Decided'

In light of these earlier results, we suspected that the effect would also arise for 'decided.' We therefore conducted an additional experiment.

Subjects were 37 undergraduate students taking philosophy classes at UNC-Chapel Hill. Each subject was randomly assigned to receive either the help vignette or the harm vignette. Subjects were then asked whether they agreed or disagreed with the statement:

- The chairman decided to help [harm] the environment.

Ratings were recorded on a scale from 1 ('disagree') to 7 ('agree'). The mean rating for the help condition was 2.7; the mean for the harm condition was 4.6. This difference is statistically significant, $t(35) = 2.4$, $p < .05$.

4. 'Advocated' and 'In Favor Of'

Given that the effect had emerged for so many other folk-psychological concepts, we predicted that we would be able to find it even if we simply selected arbitrary expressions that in some way indicated that an agent was holding or displaying a positive attitude toward a given outcome. We chose the expressions 'advocated' and 'in favor of.'

Subjects were 62 students taking undergraduate philosophy classes at UNC-Chapel Hill. The experiment used a 2x2 design, with each subject randomly assigned to receive a story with a particular moral status (harm or help) and also randomly assigned to a particular question type ('advocated' or 'in favor of').

Subjects in the harm condition received the following vignette:

> The management of a popular coffee franchise held a meeting to discuss a new procedure for preparing and serving coffee.
>
> The assistant manager spoke forcefully in favor of adopting the new procedure, saying:
>
> I know that this new procedure will mean more work for the employees, which will make them very unhappy. But that is not what we should be concerned about. The new procedure will increase profits, and that should be our goal.

Subjects in the help condition received a vignette that was almost exactly the same, except that the assistant manager argued for a policy that would mean *less* work for the employees:

> I know that this new procedure will mean less work for the employees, which will make them very happy. But that is not what we should be concerned about. The new procedure will increase profits, and that should be our goal.

Subjects were then asked whether they agreed or disagreed with a particular statement about the vignette. Each subject was randomly assigned to receive either a statement claiming that the agent 'advocated' bringing about an effect or that the agent was 'in favor of' bringing about an effect. Hence, the possible statements were:

- The assistant manager advocated [was in favor of] making the employees do more work.

- The assistant manager advocated [was in favor of] making the employees do less work.

Subjects rated each statement on a scale from 1 ('disagree') to 7 ('agree'). The results are displayed in Table 1.

	Harm	Help
Advocated	4.1	2.8
In Favor Of	3.8	2.6

Overall, there was a significant main effect such that subjects were more inclined to agree in the harm condition than in the help condition, F (1, 58) = 4.6, p < .05. There was no significant difference between the two question types ('advocated' vs. 'in favor of'), nor was there any significant interaction between moral status and question type.

Discussion

In light of these results, we are inclined to think that the impact of moral judgment is pervasive, playing a role in the application of *every* concept that involves holding or displaying a positive attitude toward an outcome. That is, for all concepts of this basic type, we suspect that there is a psychological process that makes people more willing to apply the concept in cases of morally bad side-effects and less willing to apply the concept in cases of morally good side-effects.[2]

On this view, there is nothing special about the concept of intentional action in particular that makes moral judgments relevant to it. Rather, the significance of the concept of intentional action is simply that people fall somewhere near the midpoint in their willingness to apply this concept in cases of morally neutral side-effects (Mele and Cushman 2007). Hence, when people become less willing in the morally good case and more willing in the morally bad case, their answers end up falling on opposite sides of the midpoint, and the asymmetry is therefore especially easy to detect.

If we now turn away from the topic of intentional action and begin looking at other concepts, we may find that the impact of moral considerations is not sufficient to move people's intuitions all the way from disagreement to agreement. But one should not therefore conclude that people's intuitions about these other concepts are entirely unaffected by moral considerations. On the contrary: moral considerations appear to be having exactly the same sort of impact on these other concepts that they do on intentional action; it's just that this impact is sometimes a little bit more difficult to detect.

A Tentative Hypothesis

Thus far, we have been providing evidence for the view that moral considerations affect the application of a wide array of different concepts. The question now is why so many different concepts should be subject to this same basic effect.

In addressing this question, we will be adopting a somewhat unusual approach. We will not offer anything like a full picture of any of the concepts under discussion here. Instead, we will focus only on a common element that we believe they all share.

[2] An anonymous reviewer points out that the effect might even arise for attributions of belief (e.g., for attributions of the belief that harming the environment would be a good thing). The results reported above do not speak to this issue either way, but a proper experimental study of the question could shed valuable light on the phenomena under investigation here.

The hypothesis we present here describes just this one element and does not make any claims about the other aspects of the relevant concepts.

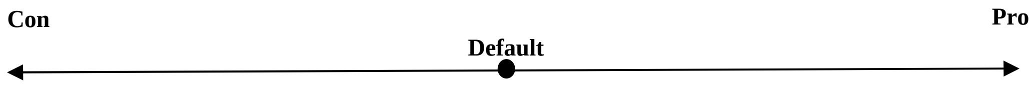
Although this approach may be disappointing to some readers, our decision to adopt it reflects a more general view about how best to make progress in this domain. Looking on the surface, one finds that people have intuitions involving various different concepts – intuitions about intentional action, intuitions about desire, and so forth. One may then be tempted to develop a separate theory about each of these types of intuitions – a theory of intentional action intuitions, a theory of desire intuitions, etc. It seems to us, however, that this view is not quite right. We find greater theoretical promise in a picture according to which each type of intuition is the product of a number of different underlying mechanisms, and each underlying mechanism contributes to a number of different types of intuition. The real aim, then, is not so much to develop theories about the various types of intuitions as to develop theories about the underlying mechanisms. On this approach, one never really ends up with anything that could be called a 'theory of intentional action intuitions.' Instead, predictions about the intuitions one finds on the surface simply fall out of one's understanding of these various mechanisms and their interaction.

Let us begin, then, by asking what sorts of psychological mechanisms might be affecting the application of all of the different concepts we have been investigating thus far. It seems to us that the common element that all of these concepts share is that each of them involves the idea of some kind of *pro-attitude* about an outcome – the idea of supporting or approaching or favoring an outcome. We suspect, then, that although a proper understanding of each of these concepts would involve a wide variety of seemingly unrelated notions, all of the concepts rely on a mechanism that distinguishes 'pro' from 'con.' It is this underlying mechanism that we propose to investigate here.

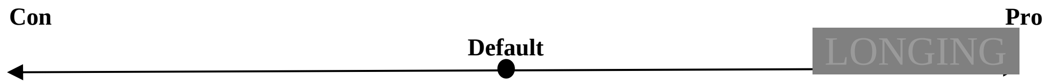
The first question to address is how people represent pro-attitudes in general. Our hypothesis is that such attitudes are represented, not in terms of a simple dichotomy between 'con' and 'pro,' but as a matter of *degree*.[3] Accordingly, the attitude an agent takes toward an outcome can be thought of as represented on a kind of scale. At one end

[3] Here we abandon certain aspects of the theory in Knobe (2006) in light of the powerful objections leveled against that theory by Malle (2007).

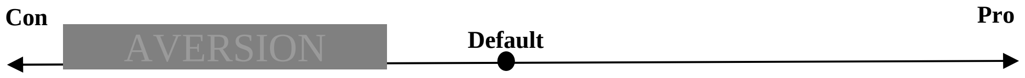
of the scale would be the state of an agent who has an overwhelmingly unfavorable attitude toward the outcome. At the other end would be the state of an agent who has an overwhelmingly favorable attitude toward the outcome. Intermediate cases would be represented by points toward the middle of the scale.



We can then suppose that different concepts require the agent to occupy different positions along this scale. Thus, the concept *desperate longing* might be represented as requiring a position very far to the 'pro' side:



By contrast, the concept *mild aversion* might require a position slightly toward the 'con' side:



Now, when we represent these different concepts using the same basic type of diagram, it is not because we think that there is literally a single type of thing – 'having a pro-attitude' – that is simply present to varying degrees in desire, intention, being in favor, and so forth. Nor are we claiming that people actually use the very same scale to understand all of the concepts under discussion here. All we mean to suggest is that all of these concepts have the same sort of underlying structure.

To get a sense of what we have in mind here, consider the semantics of the adjectives ‘interesting,’ ‘expensive,’ ‘prevalent,’ ‘amusing’ and ‘democratic.’ It certainly does not seem that there is a single unified scale underlying the semantics of each of these terms. (It would be a bit nonsensical to use a sentence like: ‘George is exactly as interesting as hamburgers are expensive.’) Still, it does appear that the semantics of all of these terms involve a similar sort of structure. All of them involve a scale from less to more (‘less interesting to more interesting,’ ‘less expensive to more expensive); all permit modification by intensifiers like ‘very’ (‘very interesting,’ ‘very expensive); all can be used with explicit comparison classes (‘pretty interesting, at least for a professor,’ ‘pretty expensive, at least for a t-shirt). In light of all these similarities, it is only natural to begin developing a very general theory that abstracts away from all the differences between these different adjectives and simply characterizes the structure that they all share (e.g., Kennedy 1999).

Our suggestion is that an analogous approach might be applicable to the concepts under discussion here. Obviously, the concepts *desiring, intending* and *in favor* differ in numerous respects, but it seems that these different concepts might nonetheless be characterized by a common structure. All of them can be understood in terms of an underlying scale that goes from ‘con’ to ‘pro’ (though the precise sense in which an attitude counts as ‘pro’ might differ considerably as one goes from one concept to the next), and all of them work by picking out points along such a scale. The goal now is to develop a theory that abstracts away from the differences among all these distinct concepts and simply describes the basic structure that they all share. Such a theory might not tell us anything about the difference between intending and being in favor, but it would tell us something very general about the patterns that arise whenever one takes a scale from ‘con’ to ‘pro’ and then constructs a concept that involves picking out certain points along this scale.

To begin with, such a theory could easily allow us to make sense of the idea that people’s responses to sentences applying folk-psychological concepts are not simply ‘yes’ or ‘no’ but instead exhibit various different levels of agreement. The basic idea would be that people show ever more disagreement as the state of the agent on the scale grows ever farther away from the state required by the concept. Thus, suppose that the concept *desperate longing* requires a position very far toward the 'pro' side. If the agent's actual attitude is extremely positive but not quite positive enough, people will feel that classifying the agent as desperately longing is *almost right* – that it is not quite true that he is desperately longing but that he still comes fairly close.

Most importantly, it might now be possible to explain how moral judgments could have a pervasive impact on people's application of folk-psychological concepts. The key notion here is that a wide variety of different folk-psychological concepts can all be characterized in terms of the same sort of underlying structure. Hence, there is no need to suppose that moral features actually figure in the representations associated with each of these individual concepts. All one needs is a single very abstract fact, namely, *that moral judgments can shift people's representations of an attitude along a scale from 'con' to 'pro.'* It then follows automatically that these judgments will come to have an impact on people's application of a huge variety of different concepts.

Further Specifying the Hypothesis

If one accepts this basic framework, a question immediately arises as to *why* people's moral judgments would affect their underlying representations of an agent's attitude. We will propose one possible answer to this question, but we first want to emphasize that the basic framework we have been developing thus far does not depend in any way on the assumption that this answer is correct. That is to say, even if the specific hypothesis we offer in the paragraphs below turns out to be entirely mistaken, there might be no need to reject the basic idea behind the approach offered here – that the pervasive impact of moral judgment arises because people's moral judgments affect all concepts that are structured in terms of a certain kind of scale.

That being said, let us now forge ahead and propose a specific hypothesis. Perhaps it will be helpful to begin with a simple analogy. Suppose that we handed out cups of coffee and cups of beer, and that we then asked people to rate the liquids in these cups as 'cold,' 'warm' or 'hot.' If the coffee and the beer were both boiling, people would presumably rate both as 'hot.' Conversely, if the coffee and the beer were both freezing, people would rate both as ‘cold.’ But now suppose that both the coffee and the beer were exactly room temperature. We might then find that people rated the coffee as ‘cold’ and the beer as ‘warm,’ even though the two liquids were in fact at precisely the same temperature.

What exactly is going on in this case? It seems that people are rating each liquid relative to a *default* that specifies what it is supposed to be like. Coffee is supposed to be at a higher temperature, beer at a lower temperature. Hence, when both are at room temperature, the coffee falls below the default (and is classified as ‘cold’), while the beer falls above the default (and is therefore classified as ‘warm’).

Now, what we want to suggest is that much the same process is at work in the phenomena we have been exploring here. Pro-attitudes are assessed relative to a default, and this default is based in part on a sense of how things are supposed to be. The key claim then is that people’s sense of what sort of attitude an agent is ‘supposed to’ have toward a given outcome can depend on the nature of the outcome itself. People are supposed to have more positive attitudes toward good outcomes, more negative attitudes toward bad ones. Hence, agents’ attitudes toward these different outcomes end up getting compared to different defaults.

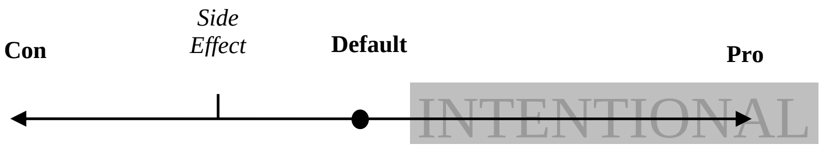
With this framework in place, we can now derive specific predictions about the intuitions people will have in different cases. The guiding assumption will be that people’s application of the word ‘intentional’ to harming vs. helping follows more or less the same pattern we saw for people’s application of the word ‘warm’ to coffee vs. beer.

First, consider a behavior that is paradigmatically intentional. The agent specifically wants to have a particular effect on the environment, and everything proceeds exactly as planned. In such a case, the agent’s attitude will be toward the ‘pro’ side of any reasonable default. Regardless of whether the act involves harming or helping, it will be classified as intentional.

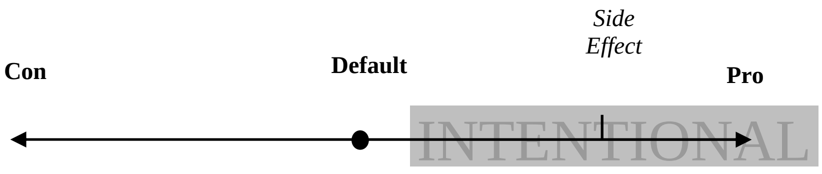
Now consider a behavior that is paradigmatically unintentional. The agent specifically wants to have no impact at all on the environment and goes out of his way to avoid having such an effect, but his plans go awry and he ends up impacting the environment anyway. Here the agent’s attitude will be toward the ‘con’ side of any reasonable default. Regardless of whether the act involves harming or helping, it will be classified as unintentional.

The thing to focus on, then, is the intermediate case. Suppose that the agent does not particularly want to impact the environment per se, but he does want to implement a program that he knows will end up having such an impact. In such a case, it may happen that the agent's attitude looks very different depending on where the default is set. In the help condition, one is inclined to think: 'How callous! Surely, any reasonable person would be at least a little bit more in favor of this outcome.' But in the harm condition, one has exactly the opposite reaction: 'How blasé! It seems like anyone should be at least a little bit more opposed to this outcome.' Hence, this very same attitude ends up falling on the 'con' side of the default in the help condition but on the 'pro' side of the default in the harm condition.

If people's moral judgments do end up shifting the default in this way, we should expect to find an effect of moral judgment on the application of certain concepts. For suppose that people represent the concept *intentionally* as requiring a position at least a little bit toward the 'pro' side of the default. Then, in the help condition, it may happen that people's attitude falls on the 'con' side of the default and that the behavior is therefore classified as unintentional:



Meanwhile, in the harm condition, that very same attitude may fall on the 'pro' side of the default, leading the behavior to be classified as intentional:



Notice now that the explanation we have offered here does not rely on any features that are peculiar to the concept of intentional action in particular. A parallel explanation could be offered for each of the other concepts discussed above: *intending*, *desire*, *in favor of*, and so on.

Testing the Hypothesis

Let us now be frank. Numerous hypotheses have already been offered to explain these phenomena, and each in turn has fallen prey to experimental falsification. It therefore seems overwhelmingly likely that our own hypothesis will turn out to be false as well. Perhaps the model adopted here will help to guide some future research, but it seems to us that we have good reason to expect that, sooner or later, some clever researcher will conduct an experiment that refutes our view.

Oddly enough, we think there might actually be good grounds for an even more specific prediction. Not only can we predict that our hypothesis will be refuted, we can predict roughly *how* it will be refuted. In each of the earlier cases, someone offered a hypothesis that predicted that the effect would only arise in a fairly circumscribed range of cases, but the hypothesis was then falsified when further research showed that the effect also arises in cases that fall outside of that range. We suspect that our own hypothesis will fall victim to the same difficulty. Sooner or later, someone is going to show that the effects we have been investigating thus far are really just one special case of a far more pervasive phenomenon.

In the meantime, though, it can be seen that the hypothesis we put forward above predicts an impact of moral judgments that goes considerably beyond the cases discussed in the previous section. In other words, the hypothesis was developed to explain the impact of moral judgments on attributions of positive attitudes in cases of side-effects, but one can see immediately that the hypothesis ends up predicting an impact of moral judgments on many other kinds of cases as well. We therefore conducted two additional experiments to examine the impact of moral considerations in cases of a rather different sort.

5. 'Opposed'

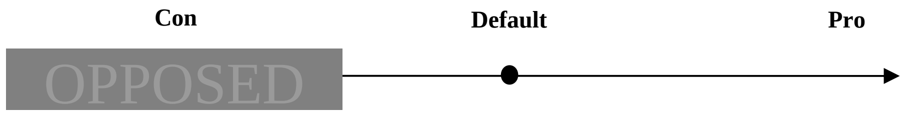
Thus far, we have been concerned exclusively with the attribution of positive attitudes: 'intending,' 'desiring,' 'in favor of,' and so forth. In each of these cases, one finds an attitude whereby the agent is favorably disposed to an outcome or motivated to pursue it. But suppose we now try to extend our investigation to negative attitudes. For example, instead of simply considering intuitions about whether an agent is 'in favor' of a given outcome, suppose we consider intuitions about whether the agent is 'opposed' to an outcome.

It follows from the hypothesis we advanced above that people's moral judgments will have an impact here too – but that this time the impact will go in the opposite direction. While people were more inclined to say that an agent was 'in favor' of harming the environment than helping it, they should be more inclined to say that an agent is 'opposed' to helping the environment than to harming it.

To see why this is so, one need only suppose that the concept *opposed* requires some negative value along an underlying scale.[4] Then people's representation of the scale

[4] Indeed, there is independent syntactic evidence that the concept involves this sort of negativity. If one looks at adjectives that involve a negative value on a scale, one finds that these adjectives license expressions like 'ever' in a way that other adjectives do not. (Thus, one can say 'It is difficult ever to admit that' but not *'It is easy ever to admit that.') One then sees the same pattern with words like 'opposed' and 'dislike.' So one can say 'He was opposed to ever doing it' but not *'He was in favor of ever doing it,' and one can say 'He disliked ever doing it' but not *'He liked ever doing it.' This syntactic phenomenon provides at least some evidence for a kind of structural similarity in the concepts picked out by these different words.

might look something like this:



Now suppose that there is a general effect whereby people shift the representation of an attitude over to the 'pro' side when they perceive the outcome as morally bad. Such an effect would have very different impacts on people's use of the phrases 'in favor' and 'opposed.' Specifically, when people determined that an outcome was morally bad, they should become *more* inclined to classify an agent as 'in favor' of it but also *less* inclined to classify an agent as 'opposed' to it.

To test this hypothesis, we conducted an additional experiment. Subjects were 56 students taking philosophy classes at UNC-Chapel Hill. Each subject was randomly assigned either to the 'harm' condition or the 'help' condition.

Subjects in the harm condition received the following vignette:

> The CEO of a company was talking with his assistant. The assistant said: 'We have conducted an in-depth study of the company's proposed new policy. Our study shows that the new policy would decrease profits for the company and that it would also harm the environment.'
>
> The CEO said: 'Look, I don't really care about what happens to the environment. What I care about is making sure that our profits don't decrease. So, with that in mind, let's definitely not implement that new policy.'

Subjects in the help condition received a vignette that was exactly the same, except that the word 'harm' was replaced with 'help.' Thus, the vignette in this condition told the story of a policy that would help the environment and an executive who was against that policy because he knew that it would decrease profits.

After reading their vignettes, subjects were asked whether they agreed or disagreed with the statements:

- The CEO was opposed to harming [helping] the environment.
- The CEO deserves *blame* for what he did.

All statements were rated on a scale from 1 to 7.

There was no significant difference between conditions on the statement about blameworthiness. (Mean for opposed to helping: 3.6; mean for opposed to harming: 2.7; *t* (54) = 1.7, *p* > .08.)

For the statement about being ‘opposed,’ ratings for subjects in the harm condition, *M* = 2.3, were significantly lower than ratings for subjects in the help condition, *M* = 3.4, *t* (54) = 2.0, *p* < .05.

6. Trying without Foresight

The hypothesis we have been developing thus far predicts an impact of moral judgment on just one particular type of representation – an underlying representation of pro-attitude. A question arises, however, as to whether one might find similar effects for other types of representation as well.

Suppose, for example, that we turn away from representations of positivity and look instead at representations of *credence*. We will be turning, then, to representations of the degree to which an agent regards an outcome as likely or unlikely. Here too, one might posit a continuous dimension. At one extreme would be the state of an agent who is absolutely certain that an outcome will occur. At the other extreme would be the state of an agent who is completely convinced that it will not. The various intermediate positions could represent different degrees of uncertainty.

What we want to know now is whether people’s moral judgments affect their representation of an agent along this continuum. We can proceed here using out usual method. First, we need to find a concept that can be correctly applied only to agents who show a particular level of credence. Then we can ask whether intermediate cases in the application of this concept are treated differently depending on their moral status.

Consider in this light the concept *intending*. If a person wishes that she could bring about an outcome but thinks that there is absolutely no chance she will succeed, one would not normally say that she 'intends' to bring about this outcome. Conversely, if she wishes to bring about an outcome and is completely convinced that she will succeed, it may begin to seem obvious that she does 'intend.' What happens then in the intermediate case? Suppose she thinks it is unlikely that she will succeed but nonetheless believes that there is at least some chance that she will bring about the outcome. The burning question is whether or not people's moral judgments will affect their application of the concept of intending in cases in this intermediate category.

To test address this question, we need to turn away from cases of side-effects and focus instead on a different kind of intermediate case. We therefore ran one final experiment. Subjects were 37 undergraduates taking philosophy classes at UNC- Chapel Hill. Each subject was randomly assigned to either the 'help condition' or the 'harm condition.' Subjects in the help condition read the following vignette:

> A man wants to defuse a bomb that will kill thousands of innocent tourists if it explodes. The only way to defuse the bomb is to enter the correct code on a keypad, but the man does not know the code. There is only a one in ten million chance of his guessing the code.
>
> The man is fully aware that there is virtually no chance that he will successfully defuse the bomb, but he desperately wants to save the tourists. So, without even looking at the keypad, he just randomly presses some keys.

Subjects in the harm condition read a vignette that was exactly the same, except that the word 'defuse' was replaced with 'detonate.' Thus, the vignette in the harm condition involved a man who is trying to detonate a bomb that will kill thousands of innocent tourists but who believes that he will probably fail.

After reading their vignettes, subjects were asked whether they agreed or disagreed with the statement:

- The man *intended* to defuse [detonate] the bomb.

To help subjects understand precisely what we were getting at here, we also included a second statement, namely:

- The man *wanted* to defuse [detonate] the bomb.

The order of these two statements was counterbalanced.

There were no significant order effects and no significant differences on the question about whether the agent 'wanted.' (Mean for harm: 6.5, mean for help: 6.6, t (35) = .21, $p > .8$.)

On the question about whether the agent 'intended,' the mean rating for subjects in the help condition was 3.8; the mean rating for subjects in the harm condition was 5.6. This difference is statistically significant, $t(35) = 2.5$, $p < .05$.

The one sad thing about this result is that it is in no way predicted by the hypothesis we presented above. That hypothesis predicts an effect in cases that are intermediate in pro-attitude but does not also predict an effect in cases that are intermediate in credence level. (Of course, the hypothesis does not specifically predict that one will *not* find an effect in this latter type of case.)

Now, one possible reaction at this point would be to suggest that there is a process whereby moral judgments affect representations of pro-attitude and then another, completely separate process whereby moral judgments affect representations of credence. But we hope that someone will be able to do better than that. Our hope is that someone will be able to find a single underlying process that explains all of the phenomena we have been discussing here. Perhaps the psychological processes we described above would then be seen as just one special case of a far more pervasive phenomenon.

Discussion

In an earlier section, we showed an influence of moral considerations in intuitions about positive attitudes toward side-effects. It now appears that the effect is not limited to that narrow range of cases. Far from it: the data presented here suggests that that effect actually arises both in cases that do not involve pro-attitudes and in cases that do not involve side-effects.

At this point, we think there is little hope of developing a theory that somehow treats each type of case individually. What is needed is a general theory about the impact of moral judgment on folk psychology. It should then be possible to show how that general theory can explain the specific types of judgments we find in these various types of cases. Given the general theory, and given some general facts about how people represent pro-attitudes and side-effects, it should be possible to derive the prediction that people's attribution of positive attitudes in side-effect cases will depend in part on their moral judgments.

Although the hypothesis proposed here is probably mistaken in certain details, it at least has the right form. The theory posits a role for moral considerations in absolutely all attributions of pro- or con-attitudes. It then predicts that moral considerations will not be sufficient to shift people's intuitions in many types of cases but that they will be sufficient to shift people's judgments in cases that have an 'intermediate' character. In other words, the framework introduced above makes it possible to derive the effects observed for positive attitudes toward side-effects from a perfectly general theory.

Conclusion

When experimental studies first began showing that moral considerations could influence the application of folk-psychological concepts, it might have been thought that this effect would be limited to a tightly constrained range of cases. One could have supposed, e.g., that the effect would only arise for the concept of intentional action, or that it would only arise in cases of side-effects, or that there would be some other, fairly narrow range of circumstances in which it could be found. It could then have been supposed that there was a kind of 'core' of folk-psychology that was entirely free of the impact of moral judgment.

Plausible though it may have seemed, this view appears not to be correct. On the contrary, as we learn more and more about the application of various different folk-psychological concepts, we are coming to find an impact of moral considerations in more and more places. It seems to us that there is now good reason to believe there are no concepts anywhere in folk psychology that enable one to describe an agent's attitudes in a way that is entirely independent of moral considerations. The impact of moral judgments, we suspect, is utterly pervasive.

References

Alicke, M. forthcoming: Blaming badly. *Journal of Cognition and Culture*.

Cushman, F. 2007: The effect of moral judgment on causal and intentional attribution: What we say, or how we think? Unpublished manuscript. Harvard University.

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

Kennedy, C. 1999: *Projecting the adjective: The syntax and semantics of gradability and comparison*. New York: Garland Publishing, Inc.

Knobe, J. 2003: Intentional action and side effects in ordinary language. *Analysis*, 63, 190-193.

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

Knobe, J. 2005: Folk psychology and folk morality: response to critics. *Journal of Theoretical and Philosophical Psychology*, 24, 252-258.

Knobe, J. 2006: The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, 130, 203-231

Knobe, J. 2007: Reason explanation in folk psychology. *Midwest Studies in Philosophy*, 31, 90-106.

Machery, E. 2008: Understanding the folk concept of intentional action: philosophical and experimental issues. *Mind and Language*, 23, 165-189.

Malle, B. 2006: Intentionality, morality, and their relationship in human judgment. *Journal of Cognition and Culture*, 6, 87-113.

Malle, B. F. 2007: The puzzle of intentionality and moral cognition. Paper presented at the Society of Personality and Social Psychology Annual Convention, Memphis, TN.

Mallon, R. 2008: Knobe vs. Machery: testing the trade-off hypothesis. *Mind and Language*, 23, 247-255.

McCann, H. 2005: Intentional action and intending: recent empirical studies. *Philosophical Psychology*, 18, 737-748.

Mele, A. and Cushman, F. 2007: Intentional action, folk judgments, and stories: sorting things out. *Midwest Studies in Philosophy*, 31, 184-201.

Nadelhoffer, T. 2004: Blame, badness, and intentional action: a reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology*, 24, 259-269.

Nadelhoffer, T. 2006: Bad acts, blameworthy agents, and intentional actions: some problems for jury impartiality. Philosophical Explorations, 9(2), 203-220.

Nado, J. 2008: Effects of moral cognition on judgments of intentionality. *British Journal for the Philosophy of Science*, 59, 709-731.

Nichols, S. and Ulatowski, J. 2007: Intuitions and individual eifferences: the Knobe effect revisited. *Mind and Language*, 22, 346-365.

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

Phelan, M. and Sarkissian, H. forthcoming: Is the 'trade-off hypothesis' worth trading for? *Mind and Language*.

Tannenbaum, D., Ditto, P.H., and Pizarro, D.A. 2007: Different moral values produce different judgments of intentional action. Unpublished manuscript. University of California-Irvine.

Wright, J. and Bengson, J. 2007: Asymmetries in folk judgments of responsibility and intentional action. Unpublished manuscript. University of Wyoming.