Action trees and moral judgment

Joshua Knobe
Yale University

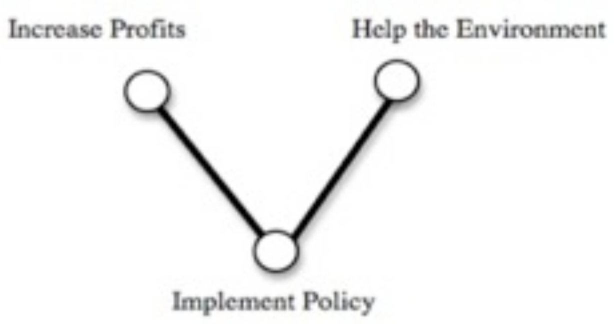
[Forthcoming in *Topics in Cognitive Science*]

It has sometimes been suggested that people represent the structure of action in terms of an *action tree*. A question now arises about the relationship between this action tree representation and people's moral judgments. A natural hypothesis would be that people first construct a representation of the action tree and then go on to use this representation in making moral judgments. The present paper argues for a more complex view. Specifically, the paper reports a series of experimental studies that appear to show that people's moral judgments can actually impact their representations of the action tree itself.

One of the most exciting theoretical developments in the study of moral cognition has been the idea that people might be representing the relationships between actions in terms of an *action tree* (Goldman 1970; Mikhail 2000, 2007).[1] This approach allows us to offer a rigorous, formal account of people's ordinary understanding of human action and then to analyze the interactions between this understanding and people's moral judgments.

So, for example, suppose that a corporate executive implements a new policy and that he thereby increases profits and also helps the environment. The suggestion would be that people's representation of the relationships that obtain among these actions can be helpfully understood using a kind of tree structure:

[1] The research reported in this paper draws heavily on a framework developed by Alvin Goldman and John Mikhail, and I am greatly indebted to their work. For comments on the experiments reported here, I am grateful to Fiery Cushman (who refuted my original account of these phenomena), Joshua Greene, John Mikhail, Jonathan Phillips, Daniel Rothschild (who helped me to develop a new account after Cushman had refuted the earlier one), Constantine Sandis, Stephen Stich, and one anonymous reviewer.



Of course, people do not usually talk about human actions by making explicit reference to tree structures, but researchers have argued convincingly that we can make inferences about people's representations of action trees just by looking at their use of ordinary expressions like 'by' and 'in order to' (see especially Goldman 1970). It may therefore be possible to arrive at a fairly good understanding of the way people think about action trees, even without ever asking them about tree structures directly.

A question now arises about the relationship between action trees and moral judgment. At least initially, it might tempting to suppose that the process of making moral judgments can be divided into a number of distinct stages. First, people would construct an action tree for the case at hand; then they would use that action tree in a series of computations that eventually yield a judgment as to whether the act is permissible or impermissible. On such a view, there is a kind of one-directional relationship between action trees and moral judgment. People's representations of the structure of the action tree would affect their moral judgments, but their moral judgments would not have any effect on their representations of the structure of the action tree itself.

Although this proposal might initially appear to have the ring of truth, much of the actual empirical work in this area now points in a radically different direction. An ever-growing body of experimental studies indicates that it is a mistake to suppose that people go through an initial stage in which they simply try to figure out what happened in a given situation, followed by a subsequent stage in which they use this information to arrive at a moral judgment. Instead, it appears that just about all of the judgments that were supposed to belong to this purely descriptive 'initial stage' can actually be influenced by moral considerations. Thus, studies have shown that people's moral judgments can impact their intuitions about whether an agent acted 'intentionally' (Knobe 2006: Nadelhoffer 2006: Young et al. 2006), whether she 'caused' certain outcomes (Alicke 2000: Cushman 2010: Knobe & Fraser 2008), whether her act counts as a 'doing' or an 'allowing' (Cushman et al. 2008), whether she counts as 'deciding' (Pettit & Knobe 2008) or 'desiring' (Tannenbaum et al. 2010) or 'valuing' (Knobe & Roedder 2009). A question now arises as to whether this same approach might be fruitfully applied to the study of people's representations of action trees.

One important insight from the existing literature here is that people might follow a principle according to which, absent conflicting evidence, one should always assume that the agent's end or goal is always to achieve a good effect, rather than a bad one (Mikhail 2007). This is an intriguing suggestion, which plays a key role in one prominent account of the way people think about action trees (Mikhail 2007).

My aim here, however, is to argue for a more radical conclusion. I want to suggest that people's moral judgments actually have an impact on their representations of the fundamental geometry of the action tree itself, affecting their intuitions about the basic elements of the tree and the ways in which these elements are interconnected. In fact, as we will see in a moment, the experimental data indicate that people's moral judgments can even impact their intuitions about which distinct actions are contained within the tree.

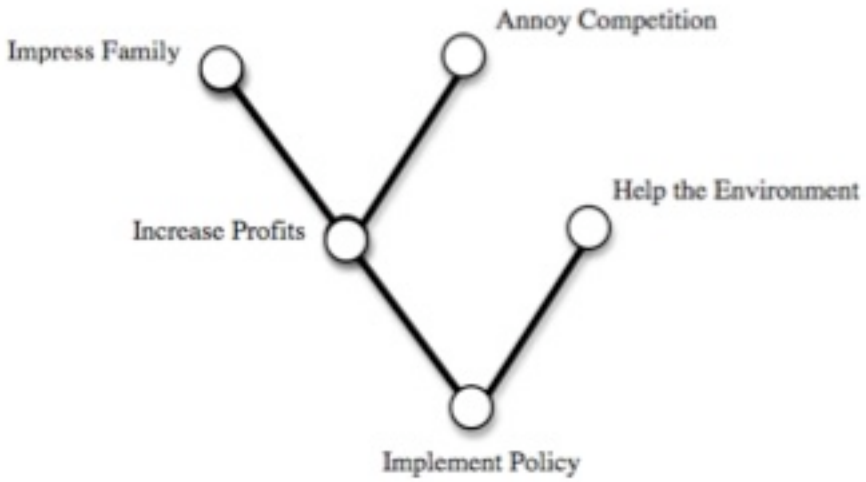
I

Thus far, we have offered only a very brief and fragmentary explanation of what action trees are supposed to be. It might be thought, therefore, that our first order of business should be to clarify some of the fundamental conceptual questions that arise here. Then, after we have a clear sense of what action trees fundamentally are, we can begin conducting experimental studies to see how people represent them.

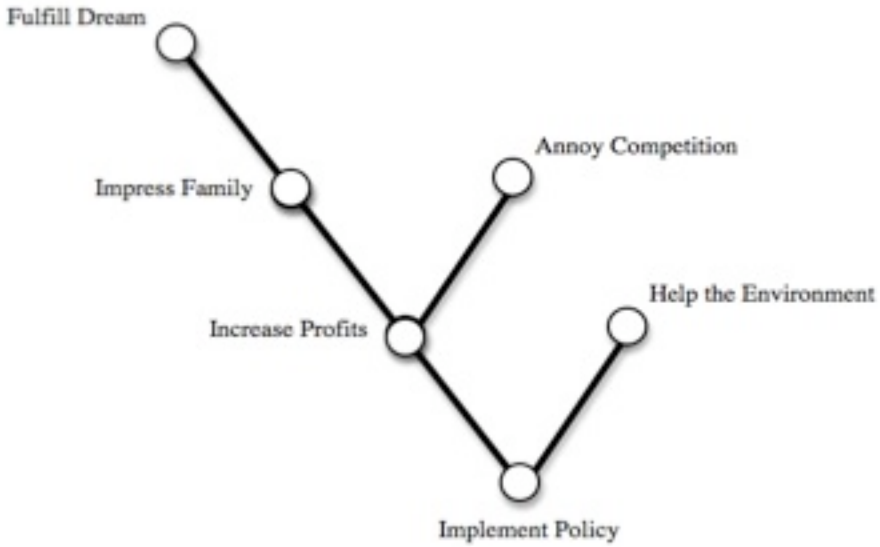
In my view, though, this gets the order of investigation exactly backwards. We don't need to *start out* with a deep understanding of the fundamental concepts at play here. This sort of deeper understanding will arrive only gradually as we continue our empirical study and reflect on the significance of the results obtained. All we need to begin with are a few vague clues that help us pick out the phenomena under investigation.

In the case at hand, we start off with two main types of clues that can guide our research. On one hand, we have some initial intuitions about what the trees themselves should look like. On the other, we have the idea that action trees are supposed to stand in a certain relationship to people's ordinary intuitions about expressions like 'by' and 'in order to.' We can now proceed by adjusting each of our views in light of the others – changing first one, then the other, until eventually we arrive at a theory that adequately explains all of the phenomena.

For our first clue, we can turn to some rough initial intuitions about how to construct the action trees themselves. The best approach here is just to introduce a few simple examples. Above, we saw an action tree for the case of an agent who increases profits and helps the environment. We can then add further elements to this story and watch as the action tree becomes more elaborate. Thus, suppose we stipulate that, by increasing profits, the chairman both impresses his family and annoys the competition. The tree would then become:

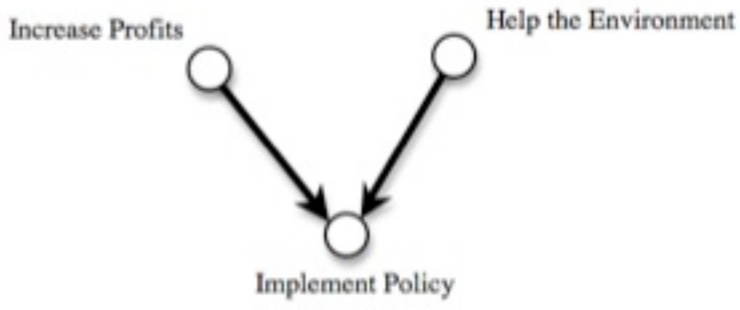


And now suppose we stipulate that, by impressing his family, he ends up fulfilling his dream. We then get an action tree like this:



These examples of action trees leave us with a kind of starting-point that can guide future research, but of course, we should be open to the possibility that we will have to give up some of the assumptions at work here in light of subsequent findings.

For our second clue, we can look to people's ordinary use of the word 'by.' As an initial hypothesis, we can suggest that people use this word to indicate movement *downward* in the action tree. In other words, the suggestion will be that people say that an agent did one thing ‘by’ doing another when one can get from the first to the second by simply moving downward in the tree. On this hypothesis, the acceptable uses of ‘by’ in our story would be the ones given in the arrows below:



It should then be acceptable to use the sentence:

- The chairman increased profits by implementing the policy.

  (which involves going straight downward)

But it should be unacceptable to say:

- The chairman implemented the policy by increasing profits.

  (which involves going upward)

And it should also be unacceptable to say:

- The chairman increased profits by helping the environment.

  (which involves going sideways)

It seems to me that intuitions about these sentences actually do conform to the predictions. So we have at least some initial evidence for our hypothesis about the relationship between action trees and intuitions about ‘by.’

If this initial hypothesis is indeed on the right track, we now have available to us a way of testing out some ideas about how people represent action trees. There is no need to start out with a full theoretical account of the nature of these trees or the meanings of their component parts. At least as a first step, we can simply present people with sentences asserting that certain things were done 'by' doing other things. Then, if we look at the patterns in people's intuitions about the acceptability of these sentences, we can begin to make inferences about the structure of the trees they are representing.

II

To explore the relationship between action trees and moral judgment, I began by conducting a simple experiment (see Appendix, Experiment 1). Subjects were randomly assigned either to the 'morally good' condition or to the 'morally bad' condition. Subjects in the morally good condition received a story about a corporate executive who increases profits and helps the environment:

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

They were then asked whether they agreed or disagreed with the sentence:

- The chairman increased profits by helping the environment.

Meanwhile, subjects in the morally bad condition received a story that was exactly the same, except that the chairman was said to have *harmed* the environment instead of helping it.

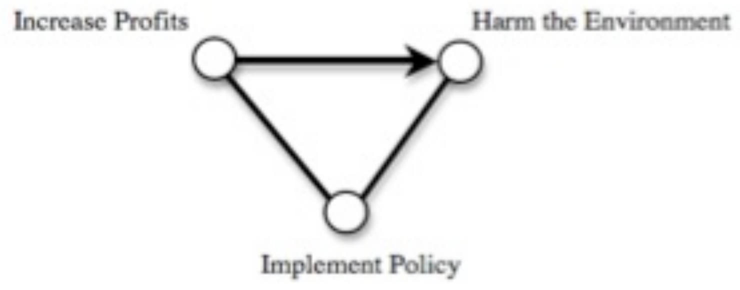
> The vice-president of a company went to the chairman of the board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.’
>
> The chairman of the board answered, ‘I don’t care at all about harming the environment. I just want to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

These subjects were then asked whether they agreed or disagreed with the sentence:

- The chairman increased profits by harming the environment.

The results revealed a surprising asymmetry. As expected, subjects in the morally good condition tended to disagree with the statement that the chairman increased profits ‘by’ helping the environment. But subjects in the morally bad condition showed a very different pattern of intuitions. The majority of them actually *agreed* that the chairman increased profits ‘by’ harming the environment. Yet it seems that the main difference between these two conditions is just that harming the environment is regarded as morally bad while helping it is regarded as morally good. So the results seem to suggest that people’s moral judgments are somehow influencing their intuitions about ‘by.’

These results are puzzling on a number of levels. First of all, there is the fact that people’s moral judgments are having an impact on their intuitions about a question like this one (which seems, at least on the surface, to call for a purely descriptive judgment). But that is not the only difficulty here. There is also something puzzling in itself about the judgment people tend to make in the morally bad case. After all, if we draw the action tree in the obvious way, we end up with the conclusion that the path from increasing profits to harming is simply going sideways:



Now, if one reflects on this action tree and on the meaning of the word ‘by,’ it can begin to seem obviously unacceptable to say: ‘The chairman increased profits by harming the environment.’ Yet the stubborn fact remains that people *do* find this sentence acceptable. In fact, I myself can’t help having the intuition that this sentence sounds right, no matter how much I think about all the reasons to conclude otherwise.

When I first encountered this problem, I didn’t know quite what to make of it, and I therefore decided to put the whole issue to one side for a while.

III

I was awoken from this slumber by an intriguing experimental result from the philosopher Joe Ulatowski. This result opened up a new way of looking at these questions and led to a series of unexpected empirical and theoretical developments.

At the time, Ulatowski was working on questions about what is sometimes called 'act individuation.' The issue here is a bit technical, but one can get a sense for the basic idea just by thinking about a classic example from the philosophical literature. Suppose that a man is operating a pump and thereby poisoning the inhabitants of a house. It seems, then, that either of the following two sentences could be a correct description of what he is doing at that moment:

‘He is operating the pump.’

‘He is poisoning the people.’

The question now is whether these two descriptions pick out two different actions or whether they are just two different ways of describing the very same action. Is the action of poisoning the people truly a separate action from the action of operating the pump, or are these really just the very same action described in two different ways?[1]

The issue here may seem like a somewhat obscure one, but Ulatowski had a hunch that he could gain a better understanding of how people thought about it just by asking them directly for their intuitions. What is more, he thought that it might be possible to show that these intuitions could actually be influenced by people’s *moral* judgments. In other words, his hypothesis was that people’s intuitions about whether two descriptions picked out the very same action might actually depend on their judgments as to whether that action was morally good or morally bad.

To test this hypothesis, Ulatowski conducted a simple and elegant experiment. Subjects in the ‘morally good’ condition received the following vignette:

> Smith’s job is to pump water into the cistern which supplies the water of a house.
>
> One day Smith operates the pump and replenishes the house’s water-supply. The occupants of the house are sick and have severe infections. Jones tells Smith that someone has found a way of systematically purifying the water’s source with a cumulative antibiotic whose effects are unnoticeable until they cure someone who has a severe infection.

> Smith says, 'I don't care about purifying the water's source; I just want to earn my pay.'
>
> The occupants of the house drink the water. Sure enough, they are saved and live.

These subjects were then asked:

- Was Smith's operating the pump the same thing as his saving the house's inhabitants or were they distinct?

Meanwhile, subjects in the 'morally bad' condition received a vignette that was almost exactly the same, except that the agent ends up having a morally bad effect on the people living in the house:

> Smith's job is to pump water into the cistern which supplies the water of a house.
>
> One day Smith operates the pump and replenishes the house's water-supply. The occupants of the house are healthy and have no health problems. Jones tells Smith that someone has found a way of systematically contaminating the water's source with a deadly cumulative poison whose effects are unnoticeable until they can no longer be cured.
>
> Smith says, 'I don't care about contaminating the water's source; I just want to earn my pay.' The occupants of the house drink the water. Sure enough, they are poisoned and die.

They were then asked the question:

- Was Smith's operating the pump the same thing as his poisoning the house's inhabitants or were they distinct?

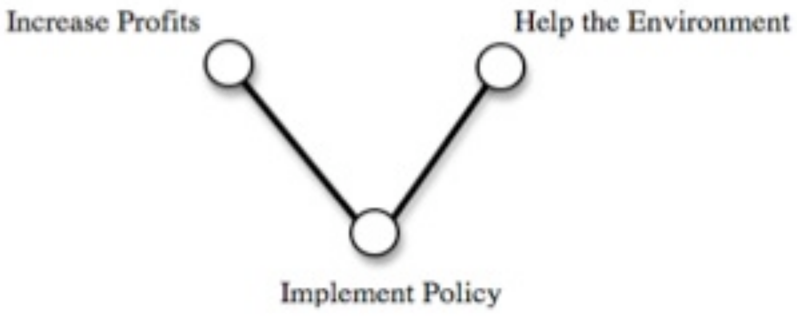
The results revealed a striking asymmetry. Subjects in the morally good condition overwhelmingly responded that the pumping and the saving were two distinct actions, but subjects in the morally bad condition had a very different response. The majority of them actually concluded that the pumping and the poisoning were *the very same action* (Ulatowski 2008).

In light of this result (and another one much like it), Ulatowski arrived at a radically new view about the way people ordinarily individuate actions. His suggestion was that people's intuitions about whether two descriptions pick out the very same act depend in a crucial way on their moral judgments. In particular, the suggestion was that people are more likely to say that the two descriptions pick out the same action when they believe that this action is itself morally bad.

IV

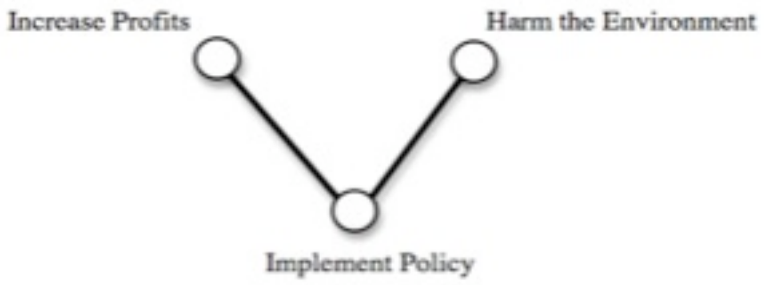
With these ideas in place, we can return to the problem we encountered about people's use of 'by.' The hope is that we will now be able to see the problem in a different light.

Recall that we had little difficulty in understanding the case in which the chairman helps the environment. There, we assumed that the action tree looked something like this:



The pattern of people's intuitions then followed naturally from the assumption that people only apply the word 'by' when going downward in the action tree.

We ran into trouble, however, when we turned to the case in which the chairman is harming the environment. There, we assumed that the action tree was represented as follows:



It then proved difficult to make sense of people's intuitions. Given that the act of increasing profits was just to the side of the act of harming the environment, it was hard to see why people were willing to say that the chairman did the former 'by' doing the latter.

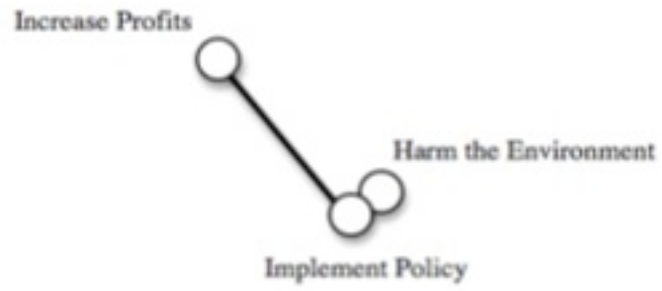
But in light of Ulatowski's work, it seems that we might have reason to rethink our assumptions about the action tree itself. The thing to focus on here is the relationship between harming the environment and implementing the policy. What we have here is a case in which there are two different ways of accurately describing what the chairman is doing at a particular time:

'He is implementing a policy.'

'He is harming the environment.'

The question now is whether these two descriptions are picking out two truly distinct actions or whether they are just two different ways of describing the very same action. Here we have good reason to make a specific prediction. People presumably regard harming the environment as morally bad, and if Ulatowski's hypothesis is to be believed, they should therefore conclude that the harming of the environment and the implementing of the policy are *the very same thing*.

We can now glimpse a path out of our difficulties. If the harming of the environment is simply the same thing as the implementing of the policy, it should not be represented on the action tree as a separate element that occupies its own distinctive position. Instead, it should be represented as occupying the same position as the implementing of the policy. So the action tree should look like this:



But notice what happens then. If the harming collapses into the implementing, one can get from the increasing to the harming just by going straight *downwards*. It now follows automatically that it should be acceptable to say: ‘The chairman increased profits by harming the environment.’

In other words, we can explain the whole pattern of people’s intuitions without introducing any special assumptions about the use of ‘by’ in particular. All we need is a very simple hypothesis about the use of ‘by’ combined with a general claim about the impact of moral judgments on representations of the action tree.

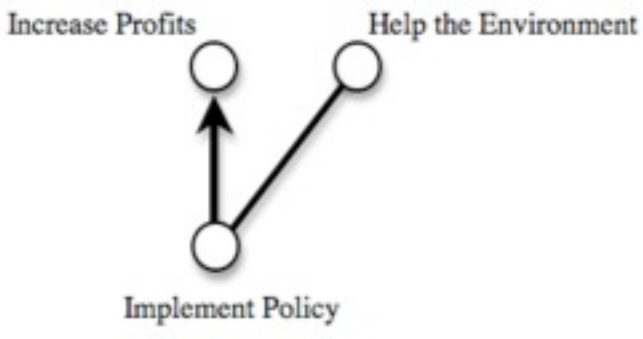
V

But in that case, it seems that we are now in the position to make a host of surprising new predictions. If moral judgments are actually impacting the representation of the underlying tree structure, the effect we observed above should not be restricted to people’s use of the word ‘by.’ On the contrary, it should arise for *any* test that allows us to tap into that tree structure. All we need to do is find some other way of getting at the same underlying representation, and we should be able to see the same sort of asymmetry arising there too.

Let us turn then to people's use of the expression *in order to*. This is the expression people use in sentences like 'He moved the table in order to make more room for the couch' or 'She emphasized her religious beliefs in order to appease Christian voters.' Many researchers believe that people's use of this expression can enable us to tap into their underlying representations of the action tree.

We can now follow a procedure almost exactly like the one we adopted above when looking at intuitions about 'by.' To begin with, we introduce a simple hypothesis about people's criteria for the use of 'in order to.' Here again, the hypothesis describes a certain kind of movement along the tree. This time, it is that people use the expression 'in order to' to indicate movement *upwards in the direction of a goal*.

Before we can apply this hypothesis, we will need to modify our representations of the action tree so that they visually represent the distinction between the agent's goal and all of the various side-effects the agent brings about along the way. We can represent the path from a behavior to its goal by using an arrow that goes straight upward while leaving all of the side-effects on the sides. Our representation of the case of helping the environment would then become:



If our hypothesis is correct, it should be acceptable to say:

- The chairman implemented the policy in order to increase profits.

  (going upwards in the direction of a goal)

But it should be unacceptable to say:

- The chairman increased profits in order to implement the policy.

  (going downwards)

And, most importantly for present purposes, it should be unacceptable to say:

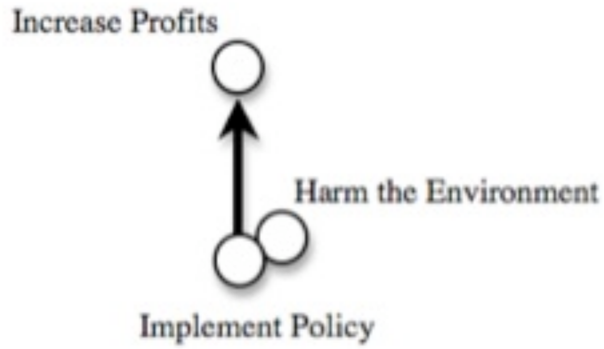
- The chairman helped the environment in order to increase profits.

  (going sideways)

As it happens, studies show that people actually do find this sentence unacceptable (Knobe 2004). They typically justify this response by saying something like: 'It wasn't really *helping the environment* that he did in order to increase profits. Rather, he implemented the policy in order to increase profits, and the helping of the environment was just a side-effect that happened to come along with this decision.'

But now suppose one changes the helping to harming. If the account we have been developing is correct, this should lead to a change in the fundamental geometry of the tree itself.

People should regard the harming as the very same thing as the implementing, and a whole branch of the tree should therefore collapse, yielding something like this:



But if the geometry of the tree changes in this way, our hypothesis predicts a corresponding change in people's intuitions about 'in order to.' Since one can now get from harming the environment to increasing profits just by going upwards in the direction of a goal, people should find it acceptable to say:

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

Remarkably enough, studies show that people *do* find this sentence acceptable (Knobe 2004). In other words, there is an asymmetry such that people think it is unacceptable to say that the agent 'helped the environment in order increase profits' but they think it is acceptable to say that the agent 'harmed the environment in order to increase profits.' (Try running the experiment on yourself – I bet you'll find that the latter sentence sounds a lot better than the former.)

When I first came across this phenomenon, I thought the only way to explain it would be to suggest that people's ordinary criteria for the use of 'in order to' were far more complex than might initially have been expected (Knobe 2007). It seems to me now that this earlier view was mistaken. We don't need any complex new hypothesis about the criteria for 'in order to.' All we need are some very simple assumptions about these criteria, combined with the independently verifiable claim that people's moral judgments have an impact on the geometry of the action tree.

VI

I cannot resist giving one more example of an effect that reveals the impact of moral judgments on the action tree. This one was uncovered by the philosopher Kristen Bell, and it has therefore come to be known affectionately as 'Bell's Inequality.'

The best way to convey the basic idea behind this last effect is just to describe Bell's original experiment (Appendix, Experiment 2). In this experiment, each subject was randomly assigned either to the 'morally good' condition or to the 'morally bad' condition. Subjects in the morally good condition received the following case:

> Imagine that you are standing in front of a button. An innocent person will be killed unless you press the button.
>
> As it happens, a bystander named Bob is betting on whether or not you will press the button. If you press the button, Bob will win $10.
>
> Please tell us whether you agree or disagree with the following statements:
>
> - You are morally obligated to press the button.
> - You are morally obligated to make Bob win the bet.

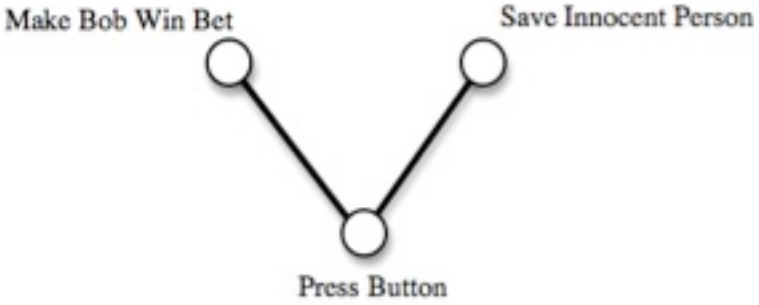
Meanwhile, subjects in the morally bad condition received a case that was almost exactly the same, except that pressing the button would lead to someone's death:

> Imagine that you are standing in front of a button. An innocent person will be killed if you press the button.

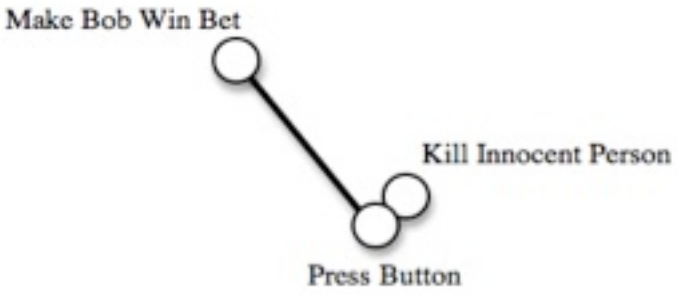
> As it happens, a bystander named Bob is betting on whether or not you will press the button. If you press the button, Bob will win $10.
>
> Please tell us whether you agree or disagree with the following statements:
>
> - It would be wrong for you to press the button.
> - It would be wrong for you to make Bob win the bet.

In this study, subjects are told about an action that clearly has a particular moral significance (saving or killing an innocent person) and are then asked whether this significance also applies to two other actions on the same tree – pressing a button and making Bob win a bet. Here again, we find an asymmetry. Subjects in both conditions ascribed moral significance to pressing the button, but when it came to making Bob win the bet, there was a pronounced difference. Subjects in the morally good condition did not agree with the claim that they were morally obligated to make Bob win the bet, but subjects in the morally bad condition actually *agreed* that it would be wrong to make him win the bet (Bell, unpublished data).

Let us now ask whether this asymmetry can be explained using our account of action trees and moral judgment. If this account is correct, subjects in the morally good condition should have generated an action tree that looks something like this:



In the morally bad condition, however, the killing of the innocent person should collapse into the pressing of the button, yielding a tree that looks more like this:



It seems natural to suppose that people think of the moral properties as applying in the first instance to the saving or the killing and then spreading from there to other actions that fall close enough to them on the tree. We can then offer at least the beginnings of an explanation for the pattern of intuitions observed in the study. The idea would be that one can only get from saving the innocent person to making Bob win the bet by going downwards and then back upwards (too far, apparently, for the moral property to spread), whereas one can get from killing the innocent person to making Bob win the bet just by going straight upwards (which, it seems, is close enough).

Of course, it would be preferable to be able to offer clear and definite criteria for the movement of moral properties on the action tree and then to show that the observed patterns follow from those criteria. I have not been able to develop criteria that satisfy this requirement, but perhaps future research will enable us to resolve these difficulties. In any case, even in the absence of clear criteria, it does seem that the present results lend support to the claim that moral judgments are impacting the geometry of the tree.

VII

With this framework in place, we can now turn to a problem that has long plagued discussions of people's ordinary understanding of action. The problem first arose in the domain of moral philosophy, and philosophers have developed increasingly complex accounts in an effort to solve it (Bennett 1995; Foot 1967; Quinn 1989). It will be seen, however, that the relevant phenomena follow quite simply from the framework developed here.

To get to the heart of this question, we can turn to an example that has been discussed at length in the existing literature on action trees – a case that is commonly known as the *trolley problem* (Foot 1967; Thomson 1985). This case comes in a number of variations, but the one we will be concerned with here goes like this:

> Ian is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Ian sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Ian is standing next to a heavy object, which he can throw onto the track in the path of the train, thereby preventing it from killing the men. The heavy object is a man, standing next to Ian with his back turned. Ian can throw the man, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Ian to throw the man? (Mikhail 2002)

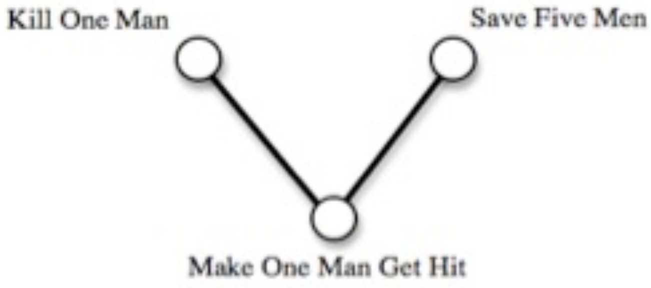
Now, suppose that Ian does decide to push the man onto the tracks. He would thereby be performing two actions – killing that one man and saving the five men walking across the tracks ahead of him. A great deal has been written about the difficult moral issues that arise in this decision, but our concern here will not be with these moral issues but rather with the seemingly straightforward question about the nature of the action tree in this case. What exactly is the relationship between the different actions the agent is performing?

The usual claim is that killing the one is a *means* to saving the five. This claim has played an important role in the philosophical literature (e.g., Foot 1967), and recent experimental work provides strong reason to believe that ordinary people construe the actions in this way as well. For example, Mikhail (unpublished data) has shown that people confronted with this case think it is right to say:

- He killed the one man in order to save the five.

It is then suggested that people's understanding of this action as a means to an end plays a key role in the moral judgments that they ultimately make about the case.

At this point, it might appear that everything is fitting together smoothly and the relationships among these actions are perfectly clear, but even in the paper in which this case was first introduced (Foot 1967), one finds an awareness that there is a deep problem here. For when we consider the case from a more theoretical point of view, it appears that the relationship among the actions should look more like this:

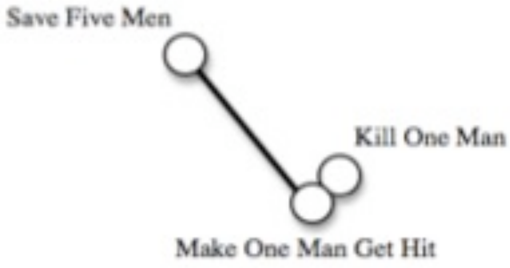


On this latter construal, the act of killing the one is not in any way a means of saving the five. Rather, the agent saves the five by *making the man get hit by a train*. Of course, if the agent makes the man get hit by a train, the agent will thereby be killing him, but that is merely an unfortunate side-effect that plays no role at all in the series of steps that lead up to the attainment of the agent’s goal. (If the man somehow emerged unscathed from his run-in with the train, the plan would still go forward without a hitch, and the five would still be saved.)

We now face a kind of conflict. Considering the case from a more theoretical perspective, it may be hard to see how we could possibly conclude that the killing of the one man was a means to the saving of the five. The obvious sorts of theoretical approaches all end up leading to the conclusion that the killing of the one is a mere side-effect. Yet when we think these questions over in a more intuitive way, we arrive at exactly the opposite conclusion. It then begins to seem ridiculous to say: ‘He wasn’t trying to kill that man; he just wanted to make sure the man got hit by a train.’ If anyone actually drew that theoretical conclusion, people might be inclined to respond by saying something like: ‘Don’t keep fussing over these technicalities! The act of making the man get hit by a train just *is* the act of killing him. There’s no need to keep worrying about the precise distinction between them.’

Now, when one first begins considering this problem, one may be tempted by the strategy of trying to resolve it by appealing to purely scientific facts about, e.g., conditional probabilities. One might start off with the thought that, given that the man is going to be hit by the train, it is simply *ridiculous* even to imagine that he would not be killed – the probability of the man surviving in such a case is almost nil. And one might then suggest that this fact about conditional probabilities leads people to ignore the distinction between the hitting and the killing and to treat the killing itself as a means of saving the five. But this sort of strategy leads immediately to obvious counterexamples. For example, given that the man is going to be hit by a train, it would also be ridiculous even to imagine that the impact will not make a sound, but people would presumably not adopt precisely the same attitude toward the making of a sound that they do toward the killing of the one man. They would not tend to say that the making of a sound was truly a means to the saving of the five. (The making of the sound is merely a side-effect that the agent happened to incur along the way.) So it seems that facts about conditional probabilities alone will not allow us to make sense of people's judgments in these cases.

In the framework under discussion here, however, it actually is possible to make theoretical sense of these intuitive judgments. Indeed, the account turns out to be surprisingly simple. Since the killing of the one man is viewed as morally wrong, our framework implies that it should collapse into the act of making him get hit by the train:



It then follows automatically that the killing of the one man is a means to the saving of the five. The tension between theory and intuition is resolved.

Notice, however, that although our solution is a quite simple one, it involves a radical departure from the traditional approach to understanding these issues. That traditional approach assumed that people first construct the whole action tree and only then set about making moral judgments. We are rejecting that assumption. Instead, we conclude that people's moral judgments are actually having an impact on their understanding of the geometry of the action tree itself.

VIII

At this point, we have a variety of different studies, using quite different types of intuitions, and they all seem to be pointing toward the same basic conclusion. It is hard to avoid the sense that there must really be something going on here. When people judge that an action is morally wrong, it seems that they truly are more inclined to conclude that certain actions which might initially appear to be distinct are, in fact, the very same thing.

A question remains, however, as to *why* people's intuitions show this peculiar pattern. Why should people's representations of the relations among actions be affected in this way by their moral judgments?

I now want to offer a hypothesis about the psychological mechanism underlying this effect. The basic idea will be that moral judgments are affecting people's intuitions about *what a person is most essentially doing* in performing a particular action. This notion is, admittedly, a rather vague and fuzzy one, but perhaps we can get a better sense for it if we look at a few examples and consider other English expressions that pick out more or less the same idea.

First, an example. Suppose that a man is working away in the kitchen. We might find that there are a number of different ways of accurately describing what he is doing: 'moving his arms,' 'disturbing some air molecules,' 'decreasing the total amount of butter in the refrigerator,' 'getting some exercise,' 'making an omelette.' Yet it seems that these different descriptions are not entirely equal. One has an immediate intuition that only the description 'making an omelette' gets at what the man is most essentially doing; all of the other descriptions just pick out various things that he happened to be accomplishing along the way. Of course, it might prove difficult to spell out precisely what it means to say that one of the descriptions is more 'essential' than the others, but I think it would be a mistake to try replacing this nebulous notion with something more clear and precise. What we really want to capture here is the rough, intuitive notion that certain descriptions do an especially good job of getting at the heart of what a person is up to in performing a particular behavior.

We can further specify the notion we have in mind by looking to other ordinary English expressions that seem to express the same basic idea. Take the expression 'just happened to end up…' It appears that this expression can be used to indicate that a particular description is highly inessential. Thus, if a man is cooking an omelettte in the normal way, one might say:

- He just happened to end up getting some exercise.

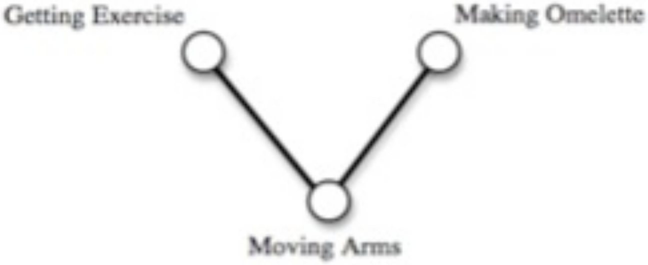
Here the words 'just happened to end up…' tell the listener that the exercise was not at all essential to the action in which he was engaged. It would sound quite odd, by contrast, to say:

- He just happened to end up making an omelette.

After all, making an omelette is not merely something that he 'just happened to end up' doing. It was the very essence of what he was up to at the time.

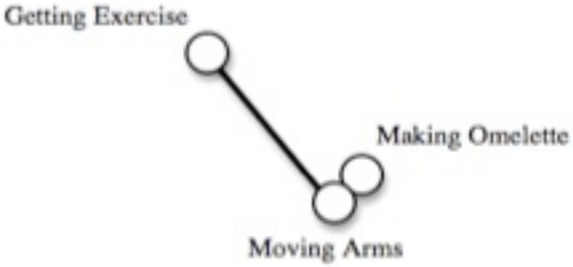
Armed with this somewhat fuzzy distinction, we can now propose a new hypothesis. The suggestion will be that people's moral judgments influence their representation of the action tree by having an effect on their intuitions about what the agent is most essentially doing in performing a particular behavior. On this hypothesis, moral considerations do not directly influence the representation of the action tree. Instead, the influence is indirect. Moral considerations affect people's intuitions about which description best gets at what the agent is most essentially doing, and there is then a general principle whereby the asymmetries we observed above arise whenever one description is seen as more essential than another.

For evidence in favor of this general principle, we can return to our example of the man working in the kitchen. Suppose that he is moving his arms and thereby both (a) getting some exercise and (b) making an omelette. In a case like this, one might initially suppose that his actions would be represented as follows:



Yet it seems that there is an important difference here between the act of making an omelette and the act of getting exercise. If the man is simply cooking in the ordinary way, one would normally conclude that the making of the omelette is at the essence of what he is doing while getting exercise is just something that happened to occur along the way.

We can now try applying our general principle. Since the making of an omelette is the very essence of what the agent is doing in moving his arms, we can predict that people will think that the arm-moving just *is* the omelette-making, and the action tree will end up looking like this:



We then predict that it should sound wrong to say:

- He made the omelette by getting some exercise.

but it should sound right to say:

- He got some exercise by making the omelette.

A simple study shows that these predictions are, in fact, borne out (Appendix, Experiment 3). So we now have at least some preliminary evidence in favor of our general principle.[2]

The next step is to show that moral considerations can actually influence people's intuitions about what is most essential in a person's actions. The claim will be that morally bad side-effects are seen as essential to the behavior (much like making an omelette, in our earlier example) whereas morally good side-effects are seen as accidental or inessential (more like getting exercise). Hence, turning back to the case we introduced at the beginning of the paper, the thought is that people see the act of harming the environment as more essential to the chairman's behavior than the act of helping is.

We can find some evidence in support of this claim by looking at people's use of the expression 'just happened to end up…' Thus, suppose someone described the chairman in the help case by saying:

- He just happened to end up helping the environment.

Assuming that people do not think that helping the environment lies at the essence of what the chairman was doing, we should predict that they would find this sentence perfectly acceptable. But now suppose someone described the chairman in the harm case by saying:

- He just happened to end up harming the environment.

If we assume that people think of the harming of the environment as lying at the very essence of what the chairman was doing, we should predict that they would have a very different reaction to this second sentence. Though they should regard the phrase 'just happened to end up' as acceptable in the help case, they should regard it as unacceptable in the harm case. A final study shows that people's intuitions do, in fact, follow this pattern (Appendix, Experiment 4).

We now have the beginnings of an explanation for the puzzling asymmetries described above. At the heart of this explanation is the idea that people regard certain descriptions as specifically getting at the essence of what a person is doing in performing a behavior. We then claim:

> (1) that there is a general principle whereby the asymmetries arise whenever one description is seen as more essential than another
>
> (2) that moral considerations have an impact on the degree to which people regard descriptions as essential.

These claims are susceptible to independent confirmation, and we have provided some evidence in support of each. Far more needs to be done before we can arrive at a fully adequate explanation of these phenomena, but what we have here is at least a tantalizing start.

IX

We now arrive at a deep and fundamental problem which I have thus far been unable to resolve. As we saw at the onset, a number of theorists have suggested that people's representations of the action tree can influence their moral judgments. This claim has received a great deal of experimental support (Greene et al. 2009; Mikhail 2007), and I think that it is probably true. Yet the studies reported here indicate that there is also an effect in the opposite direction: people's moral judgments can actually influence their representations of the action tree itself. At this point, then, we have evidence for the claim that people's moral judgments can be influenced by their representations of the action tree, but we also have evidence for the claim that people's representations of the action tree can be influenced by their moral judgments.

A key task now is to develop a theoretical model that allows us to reconcile these two claims. Just thinking about the question for a moment, one can easily come up with some plausible ways of addressing this problem. But the goal is not simply to find a solution that sounds plausible; we need a way of figuring out which of the plausible solutions is actually correct. This will be an important topic for future research.

Appendix

Experiment 1

Subjects were 46 students taking undergraduate philosophy classes at the University of North Carolina-Chapel Hill. Each subject was randomly assigned either to the 'morally good' condition or to the 'morally bad' condition. These subjects then received short vignettes and were asked whether they agreed or disagreed with certain sentences, as described in the body of the text.

Subjects rated each sentence on a scale from 1 ('disagree') to 7 ('agree'). The mean response for the morally good condition was 2.95; the mean for the morally bad condition was 4.38. This difference was statistically significant, $t(44) = 2.0, p < .05$.

Experiment 2 (conducted by Kristen Bell)

Subjects were 117 students taking undergraduate philosophy classes at the University of North Carolina-Chapel Hill. Each subject was randomly assigned either to the 'morally good' condition or to the 'morally bad' condition. These subjects then received the short vignettes described in the text and were asked whether they agreed or disagreed with certain sentences.

Subjects in the morally good condition received the sentences:

- You are morally obligated to press the button
- You are morally obligated to make Bob win the bet.

Subjects in the morally bad condition received the sentences:

- It would be wrong for you to press the button.

- It would be wrong for you to make Bob win the bet.

Subjects rated each sentence on a scale from 1 ('disagree') to 7 ('agree'). The order of sentences was fixed, with the sentence about pressing the button always preceding the sentence about making Bob win the bet.

On the sentence about pressing the button, subjects gave a slightly higher rating in the morally bad condition ($M = 6.6$) than in the morally good condition ($M = 6.0$), but this difference did not quite reach significance, $t(115) = 1.9$, $p = .06$. On the sentence about winning the bet, by contrast, ratings for the morally bad condition ($M = 6.1$) were significantly higher than those for the morally good condition, ($M = 2.2$) $t(115) = 10.4$, $p < .001$.

Experiment 3

Subjects were 32 people spending time in a New York public park. All subjects received the following vignette:

> Imagine a chef in a kitchen who is making some breakfast in the ordinary way. His arms are moving in just the way one would expect for a chef in his position. He is thereby making an omelette. He is also getting some exercise.

They were then asked whether they agreed or disagreed with the following statements:

- He got some exercise by making an omelette.
- He made an omelette by getting some exercise.
- He made an omelette by moving his arms.
- He moved his arms by making an omelette.
- He got some exercise by moving his arms.

- He moved his arms by getting some exercise.

Subjects rated each of these sentences on a scale from 1 (‘disagree’) to 7 (‘agree’). The statements were presented in counterbalanced order.

The mean rating for each statement was as follows:

5.8: Exercised by making an omelette

2.1: Made an omelette by exercising

5.9: Made an omelette by moving arms

4.0: Moved arms by making an omelette

6.3: Exercised by moving arms

2.7: Moved arms by exercising

As predicted, the rating for the statement about exercising by making an omelette ($M = 5.8$) was significantly higher than the rating for the sentence about making an omelette by exercising ($M = 2.1$), t (31) = 8.8, $p < .001$.

The study also provides a further result that might prove helpful in testing the hypothesis (given in footnote 2) that although people regard the chef’s making of the omelette as standing ‘close’ on the action tree to the moving of his arms, they do not regard these two actions as being entirely identical. Specifically, ratings for the statement about making an omelette by moving his arms ($M = 5.9$) were significantly higher than their ratings for the statement about moving his arms by making an omelette ($M = 4.0$), t (31) = 3.5, $p = .001$. This result seems clearly to indicate that there are cases in which people do draw a distinction between the two actions.

Experiment 4

Subjects were 43 people spending time in a New York public park. Each subject was randomly assigned either to the ‘morally good’ condition or to the ‘morally bad’ condition. Subjects in the morally good condition received the story about the chairman who helped the environment and were then asked whether they agreed or disagreed with the sentences:

- The chairman of the board just happened to end up helping the environment.
- The chairman of the board just happened to end up increasing profits.

Subjects in the morally bad condition received the story about the chairman who harmed the environment and were then asked whether they agreed or disagreed with the sentences:

- The chairman of the board just happened to end up increasing profits.
- The chairman of the board just happened to end up helping the environment.

Each sentence was rated on a scale from 1 (‘disagree’) to 7 (‘agree’). The order of sentences was counterbalanced.

As predicted, there was a significant effect whereby subjects tended to agree with the claim that the chairman ‘just happened to end up’ helping the environment ($M = 5.7$) but to disagree with the claim that the chairman ‘just happened to end up’ harming the environment ($M = 2.4$), $t(41) = 4.6$, $p < .001$. There was no significant difference between intuitions about whether the chairman ‘just happened to end up’ increasing profits in the morally good condition ($M = 2.6$) and the morally bad condition ($M = 3.0$), $t(41) = .4$, $p > .6$.

Notes:

[1] This question was at the heart of a philosophical debate that pitted Anscombe (1957, 1979) and Davidson (1963) against Goldman (1970). The specific example of the man pumping water and poisoning the inhabitants is due to Anscombe (1957).

Anscombe (1957, 1979) herself developed a view according to which, even if *x* and *y* are exactly the same thing, an agent might perform an action by doing *x* but not by doing *y*. The hypothesis defended in the present paper starts out with the assumption that ordinary people reject this view. To the extent that the hypothesis successfully predicts and explains people's intuitions about cases, we will have evidence that this assumption is an accurate one.

[2] For simplicity, I have been writing as though people make a dichotomous distinction between 'essential' and 'inessential,' but people's actual representations presumably involve a whole *continuum* of different degrees of essentialness. Indeed, it seems probable that the entire framework developed in the text would actually be better understood in more continuous terms. First we would replace the dichotomy between 'essential' and 'inessential' with a whole continuum of different degrees of essentialness; then we would replace the dichotomy between 'occupying completely separate nodes on the action tree' and 'completely collapsing into a single node' with a whole continuum of different degrees to which nodes can be 'close' to one another. The continuous representations of essentialness could then impact the continuous representations of closeness.

If these phenomena truly are best understood in continuous terms, we will need a slightly different account of people's intuitions about sentences like:

- He is getting some exercise by making an omelette.

It will not be quite accurate to say that the making of the omelette completely collapses into the moving of the arms, yielding a single node that permits no distinction between the two actions. Rather, the two actions are simply regarded as sufficiently 'close' to each other that one can get from the exercise to the omelette-making by going almost entirely downward in the tree.

References

Alicke, M.D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556-574.

Anscombe, G.E.M. (1957). *Intention*. Oxford: Blackwell.

Anscombe, G.E.M. (1979). Under a description. *Noûs, 13*, 219-233.

Bell, K. (unpublished data). University of North Carolina – Chapel Hill.

Bennett, J. (1995). *The Act Itself*. New York: Oxford.

Cushman, F. (2010). The effect of moral judgment on causal and intentional attribution: What we say, or how we think? Unpublished manuscript. Harvard University.

Cushman, F., Knobe, J. & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition, 108* (1), 281-289.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, *60*, 685-700.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5-15.

Goldman, A. (1970). A theory of human action. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Greene, J.D., Lindsell, D., Clarke, A.C., Nystrom, L.E., and Cohen, J.D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111* (3), 364-371.

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

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.

Knobe, J. & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong, *Moral Psychology*, Cambridge, MA: MIT Press. 441-448.

Knobe, J. & Roedder, E. (2009). The ordinary concept of valuing. *Philosophical Issues, 19* (1), 131-147.

Mikhail, J. (2000). Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'A Theory of Justice.' Ph.D. dissertation. Cornell University.

Mikhail, J. (2002). Aspects of the Theory of Moral Cognition: Investigating Intuitive Knowledge of the Prohibition of Intentional Battery and the Principle of Double Effect. Unpublished manuscript. Georgetown Law School.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences. 11*, 143-152.

Nadelhoffer, T. (2006). On trying to save the simple view. *Mind & Language. 21* (5), 565-586.

Pettit, D. and Knobe, J. (2008). The pervasive impact of moral judgment. *Mind & Language, 24* (5), 586-604.

Quinn, W. (1989). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy and Public Affairs* 18: 334–51.

Tannenbaum, D., Ditto, P.H., & Pizarro, D.A. (2010). Different moral values produce different judgments of intentional action. Unpublished manuscript. University of California-Irvine.

Thomson, J. (1985). The trolley problem. *Yale Law Journal. 94*, 1395-1415.

Ulatowski, J. (2008). How many theories of act individuation are there? PhD Thesis, University of Utah.

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.