Morality and Possibility

Joshua Knobe

[Knobe. J. (in press). Morality and possibility. In J. Doris & M. Vargas (eds.), *The Oxford Handbook of Moral Psychology*.]

Over the course of the past decade or so, there has been a great deal of research in experimental philosophy on people's judgments concerning action and agency. Some of this research has been concerned with judgments about explicitly moral questions (e.g., judgments about moral praise and blame), but much of it has fail for been concerned with judgments that might appear to be entirely non-moral. There have been numerous studies on people's judgments about whether an agent acted freely, whether an action caused some further outcome, whether an action was performed intentionally or unintentionally.

This research has revealed something surprising. As we will see in a moment, people's moral judgments appear to influence their judgments about all of these seemingly non-moral questions. In other words, people's beliefs about the moral status of an action seem to influence their judgments about whether the agent acted freely, whether the action caused further outcomes, and whether the action was performed intentionally.

Why might people's judgments about these apparently non-moral questions be influenced by moral considerations? A variety of different hypotheses immediately suggest themselves. Perhaps the effect can be explained in terms of people's emotional reactions, or perhaps it can be explained in terms of motivated cognition, or in terms of conversational pragmatics. With a little bit of further reflection, one can easily develop numerous other plausible approaches.

The present chapter focuses on just one of these approaches. On this approach, the surprising effects of moral judgment are ultimately to be understood in terms of something about the way people think about *alternative possibilities*. When you are thinking about the actual state of affairs and trying to understand what an agent actually did, you often do so by thinking about other possible states of affairs or other possible actions the agent might have performed. The core idea then is that people's moral judgments impact their judgments about the actual world because they influence the way people think about such alternative possibilities.

Explanations that invoke alternative possibilities have been proposed by a wide variety of different researchers, coming out of numerous different disciplines and theoretical backgrounds (Blanchard & Schaffer, 2013; Cova, Lantian & Boudesseul, 2016; Egré & Cova, 2015;

Falkenstien, 2013; Halpern & Hitchcock, 2015; Kominsky, Phillips, Gerstenberg, Lagnado & Knobe, 2015; Icard, Kominsky & Knobe, 2017; Kratzer, 2013; Phillips & Cushman, forthcoming; Young & Phillips, 2011). On the view I will be defending here, these different explanations should be seen as different ways of working out the details of a single basic hypothesis. Thus, it may prove helpful for many purposes to group together all of these explanations, treating them as a single approach which can then be contrasted with any of the many other approaches researchers have developed to explain these effects (such as explanations in terms of motivational bias, conversational pragmatics or mental state inference; Alicke, Rose & Bloom, 2011; Machery, 2008; Nichols & Ulatowski, 2007; Ngo, Kelly, Coutlee, Carter, Sinnott-Armstrong & Huettel, 2015; Samland & Waldmann, 2016; Sytsma, Livengood & Rose, 2012; Uttich & Lombrozo, 2010).

It should be noted, however, that the different explanations I will be grouping together might not at first appear to be very similar at all. In fact, the existing literature has been structured in such a way that these different explanations are usually treated as more or less unrelated ideas, belonging to completely separate fields.

First, the existing literature tends to be structured around the study of one or another specific type of judgment. Thus, there is a stream of papers that explore the patterns in people's judgments about freedom and then, completely separately, a stream of papers that explore people's judgments about intentional action. Papers within each of these streams tend to focus in real detail on the particular type of judgment they investigate but not to discuss questions about how the different types of judgments relate to each other.

Second, and perhaps more importantly, these explanations are often spelled out using formal frameworks, but different researchers have turned to frameworks of quite different types. Some have drawn on frameworks from linguistic semantics that make use of logic and set theory; others have drawn on frameworks from computational cognitive science that rely on probability theory.

The result is that these different strands of research end up looking almost completely unrelated. Suppose you pick up a paper that uses tools from set theory to understand the impact of moral considerations on judgments about freedom. Then, the next day, you pick up a different paper that uses tools from probability theory to understand judgments about causation. You

might find both papers interesting or helpful, but it might be a little bit hard to believe that they are really getting at more or less the same thing.

Nonetheless, that is the position I will be defending here. I argue that these different strands of research are best understood as different ways of working out the details of a single larger vision. I will refer to this larger vision as the *possibility hypothesis*.

To make this argument, it will be necessary to adopt a slightly different focus from the one that is customary within existing work in this field. Typically, research in this field is concerned with the precise patterns observed in experiments about some specific type of judgment. The present chapter will adopt the opposite strategy. I will say relatively little about the more detailed questions that have been the primary focus of most existing work. Instead, the chapter will be concerned almost entirely with the big picture, i.e., with an attempt to spell out the possibility hypothesis in a fully general way.

1. Impact of moral judgment

We begin with a very brief description of the impact of moral judgment on judgments about freedom, causation and intentional action. Although numerous studies have been conducted on each of these effects, we will be relying here on just one example for each. These examples will then prove helpful later on as we consider possible explanations.

a. Freedom

Consider first the distinction between cases in which an agent acts freely and cases in which an agent is forced by her situation to perform some action. It might seem at first that people can draw this distinction without thinking at all about the moral status of the action itself, but recent studies suggest that moral considerations actually do play a role here.

In one such study (Phillips & Knobe, 2009), participants were randomly assigned to receive one of two vignettes. Here is the vignette in which the agent ultimately performs the morally good action:

> At a certain hospital, there were very specific rules about the procedures doctors had to follow. The rules said that doctors didn't necessarily have to take the advice of consulting physicians but that they did have to follow the orders of the chief of surgery.

> One day, the chief of surgery went to a doctor and said: 'I don't care what you think about how this patient should be treated. I am ordering you to prescribe the drug Accuphine for her.'
>
> The doctor had always disliked this patient and actually didn't want her to be cured. However, the doctor knew that giving this patient Accuphine would result in an immediate recovery.
>
> Nonetheless, the doctor went ahead and prescribed Accuphine. Just as the doctor knew she would, the patient recovered immediately.

In the vignette in which the agent ultimately performs the morally bad action, the last two paragraphs become:

> The doctor really liked the patient and wanted her to recover as quickly as possible. However, the doctor knew that giving this patient Accuphine would result in her death.
>
> Nonetheless, the doctor went ahead and prescribed Accuphine. Just as the doctor knew she would, the patient died shortly thereafter.

Participants tended to say in the first case that the agent was forced to prescribe Accuphine, whereas they tended to say in the second case that the agent was not forced but rather freely chose to prescribe Accuphine (Phillips & Knobe, 2015). Further studies have shown similar effects (Phillips & Cushman, 2017; Young & Phillips, 2011), providing converging evidence that people's moral judgments actually influence their judgments about whether an agent acted freely.

1.2 Causation

Analogous effects arise for people's judgments of causation. To illustrate, consider the following vignette:

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

Now ask yourself whether it would be correct to say: 'Professor Smith caused the problem.'

Then consider a version in which the second paragraph is exactly the same, but the first paragraph is altered to suggest that the faculty member's behavior is wrong or bad.

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, **but faculty members are supposed to buy their own**. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that **only administrative assistants** are allowed to take the pens.

Here again, the question is whether it would be right to say: 'Professor Smith caused the problem.'

Strikingly, the difference in perceived wrongness between the actions in these different vignettes is reflected in a difference in causal judgment. Participants are more inclined to say that the professor *caused* the problem in the version where she is doing something wrong (Phillips, Luguri & Knobe, 2015). Numerous other studies show similar effects of moral judgment on causal judgment (e.g., Cushman, Knobe & Sinnott-Armstrong, 2008; Fraser & Knobe, 2008; Kominsky et al., 2015).

1.3 Intentional action

Finally, consider the distinction people make between actions that are performed intentionally and those that are performed unintentionally. Here too, we find a surprising effect of moral judgment.

To illustrate, here is a vignette in which the outcome is bad:

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

And here is the corresponding vignette in which the outcome is good:

> 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**.

The agent seems to have exactly the same attitude toward the outcome in these two cases, namely, complete indifference. Nonetheless, participants tend to say in the first case that the agent intentionally harmed and in the second case that he unintentionally helped (Knobe, 2003). This effect, too, has been observed in numerous studies (e.g., Ngo, Kelly, Coutlee, Carter, Sinnott-Armstrong & Huettel, 2015; Young, Cushman, Adolphs, Tranel & Hauser, 2006).

2. Morality and possibility: The core idea

At least in principle, it could certainly be the case that these three effects are due to three completely unrelated processes, but given the obvious similarities between them, it is certainly

tempting to seek a unified account that can explain all three. We will focus here on the idea that all of these effects might be explained in terms of something about the way people ordinarily think about *possibilities*.

As noted above, different researchers have spelled out this idea using quite different formal frameworks. These frameworks have enabled researchers to develop hypotheses in explicit detail, thereby generating clear testable predictions, and the result has been a great deal of valuable progress. Yet, at the same time, I worry that the use of these theoretical frameworks may have obscured one important aspect of the research conducted thus far. In particular, when different researchers are working within different frameworks, it may be difficult to see the more abstract sense in which they are actually trying to develop the same core idea.

Let us therefore start out by forsaking all theoretical frameworks and simply describe the core idea at a rough, intuitive level. The hope is that this intuitive description will make it easier to see what is shared among hypotheses that have been worked out in different frameworks and may therefore appear on the surface to be entirely unrelated.

Understood at this rough, intuitive level, the core idea has two basic elements.

2.2 Moral judgment and alternative possibilities

People are capable of considering a wide variety of alternative possibilities, but they do not seem to treat all such possibilities equally. Instead, they regard some possibilities as *relevant* and others as *irrelevant*.

For example, suppose you have just finished taking a difficult exam, and you are thinking about how things could have gone differently. One possibility you could consider would be:

> What if I had studied harder?

Another would be:

> What if the school had been destroyed by a freak natural disaster?

You might be capable of considering either of these possibilities, but all the same, it seems that there is an important difference between the two. In some sense, the first possibility seems relevant, while the second seems irrelevant. We will be turning in later sections to questions about how to spell out this idea in more detail, but for the moment, we can leave it at a rough,

intuitive level. Without relying on any specific theory, we are simply introducing the assumption that people regard some possibilities as more relevant than others.

A question now arises as to what factors influence people's judgments about the relevance of possibilities. Clearly, many different factors will play a role here. People might regard possibilities as more relevant when they are highly probable or frequent, or when they are in keeping with physical laws. Then, when it comes to possibilities involving human action, certain additional considerations seem to play a role. For example, people might regard as relevant possibilities that involve actions that are rational or that are especially good ways for an agent to achieve her goals.

Our focus here, however, will be on just one of these many factors. People's judgments about the relevance of possibilities seem to depend in part on their *moral judgments*. In particular, people seem to show a general tendency to regard possibilities as more relevant when they are morally good than when they are morally bad.

The basic idea here can be illustrated with a simple example. Suppose you believe that the electorate tends to react to outsiders with fear and hate but that it would be morally better for them to respond with compassion. On a particular occasion when the electorate responds with fear and hate, you might think:

> What if they had instead responded with compassion?

Of course, you might believe that they were highly unlikely to respond in this way, but all the same, you might regard this alternative possibility as highly relevant. You would regard it as relevant not because you thought it was especially probable but rather because you thought it was morally good.

To sum up: People regard some possibilities as relevant, others as irrelevant. One factor that influences judgments about the relevance of possibilities is moral judgment. In general, people tend to regard possibilities as especially relevant to the extent that they believe those possibilities to be morally good.

2.1 The importance of alternative possibilities

The second key idea is about the impact of regarding an alternative possibility as relevant. The idea is that people will develop quite different views about the things that actually happen depending on which alternatives they see as the relevant ones.

Again, this idea will be explored in far more detail in the sections below, but we can get a rough sense for it just by considering a simple example. Suppose that you decide to go to graduate school to study moral psychology. People might make sense of this decision by thinking about various other options you might have chosen instead. Now suppose that different people focus on different alternatives. Some people focus on the fact that you could have decided to get a more ordinary corporate job, while others focus on the fact that you could have started an indie rock band. In other words, some people are thinking:

> You chose to become a graduate student instead of taking a position where you do something far less exciting but have far better wages and job security.

While others are thinking:

> You chose to become a graduate student instead of spending your time touring the country, sleeping on people's couches and playing poorly attended shows in dimly-lit bars.

The key point now is that people's judgments about the decision you actually made will depend in large part on which of these alternatives they consider. Those who focus on the former alternative will see your decision as adventurous; those who focus on the latter might see it as stodgy or conservative.

Before continuing onward, we should note two things about this sort of effect. First, it is an effect on the way people think about *what actually happened*. For example, when people consider possibilities in which you join an indie rock band, this does not merely impact people's way of thinking about those alternative possibilities; it impacts people's way of thinking about the decision you actually made.

Second, this impact is *pervasive.* That is, it is not just an impact on certain specific judgments (say, just on judgments about adventurousness). Rather, it is an impact on people's basic way of understanding what happened, and it should therefore lead to changes in numerous

different kinds of judgments. Thus, this effect at least holds out the potential to explain the impact of moral considerations on a wide variety of different kinds of judgment.

2.3 Putting it all together

Putting these two elements together, we have the outlines of an explanation. The first element is that people's judgments about the relevance of alternative possibilities can be shaped by moral considerations. The second is that people's interpretation of what actually happened can be shaped by which alternative possibilities they regard as relevant. Together, these two elements suggest that people's interpretations of what actually happened can be shaped by moral considerations. It is this core idea that constitutes what I call the *possibility hypothesis*.

The suggestion is that the possibility hypothesis can explain each of the effects reviewed above. I return to each effect in more detail below (§3). To foreshadow, although the explanation of each effect will involve some complex further details, each explanation will include as one element the straightforward application of the core idea introduced in the present section.

[*Freedom*] People regard as relevant the possibility in which the doctor disobeys the chief of surgery to save the patient, but they regard as irrelevant the possibility in which the doctor disobeys the chief of surgery to kill the patient,.

[*Causation*] People regard as relevant the possibility in which the professor refrains from taking a pen, but they regard as irrelevant the possibility in which the administrative assistant refrains from taking a pen.

[*Intentional Action*] People regard as relevant the possibility in which the chairman actively seeks to help the environment, but they regard as irrelevant the possibility in which the chairman actively seeks to harm the environment.

Now, if we thought that these effects were due to something quite specific about moral judgments in particular, it would be natural to suppose that the key thing to focus on first would be getting some conceptual clarity on questions related specifically to moral judgment. We would want to begin by distinguishing carefully between moral judgments and other types of

judgments and, within the domain of moral judgments, between a number of different types (judgments of wrongness, judgments of blame, etc.). Then we would want to understand which specific type of judgment led to these effects. At one point, it was widely believed that these effects were indeed specific to morality, and at that point, there was a fair amount of work aimed at getting clarity on precisely these questions (e.g., Knobe 2007; Phelan & Sarkissian, 2008).

By contrast, if the present hypothesis is on the right track, the effect should not be in any way limited to morality. It should arise for moral judgments (as in the cases we have been describing here), but it should also arise for any other type of judgment that impacts the degree to which people regard certain possibilities as relevant. Existing research provides support for this prediction, indicating that these effects arise for moral judgments but also for probability judgments, judgments of rationality and judgments about purely conventional norms (e.g., Icard et al., 2017; Phillips & Cushman, 2017; Proft, Dieball, & Rakoczy, in press).

Thus, if this hypothesis is correct, the explanation requires conceptual clarity in a somewhat different place. Where we most need clarity is not so much in our understanding of moral judgment specifically as in our understanding of the ways in which people represent the relevance of possibilities.

3. Formal frameworks

Thus far, we have been invoking the somewhat nebulous notion that people regard certain possibilities as 'relevant' and others as 'irrelevant.' Within the existing literature, however, most research does not simply rely on this nebulous notion. Instead, researchers aim to spell out more precisely how people distinguish between different kinds of possibilities. This section briefly reviews three theoretical frameworks that have been used to spell out this distinction.

These three frameworks come out of different intellectual traditions, make use of different formal machinery, and are widely regarded as completely unrelated ideas. I will argue that this view is a mistaken one. A more accurate view would be that the three frameworks are best understood as three attempts to characterize a single underlying phenomenon.

Within existing work, all three of these frameworks are usually discussed using formal mathematical techniques. In this brief review, I will instead be describing them using ordinary English. Readers who are dissatisfied with this description can turn to the papers cited within each subsection for more mathematical detail.

a. Modality

Within natural language, people often talk about possibility by using expressions like 'can,' 'must,' and 'have to.' These expressions are known as *modals*. Thus, one obvious way to capture people's ordinary understanding of possibility is to look to the frameworks that have been introduced within work in formal semantics on natural language modals.

One striking fact about modals is that people seem to use them for a variety of different purposes. For example, (1a) seems to be saying something about physical laws, (1b) about moral obligations, and (1c) about ways to attain one's goals.

(1) a. No particle can go faster than the speed of light.
b. You can't keep doing that to her – you have to start treating her like a human being.
c. If you want to get to Harlem, you can just take the A train.

It might therefore appear at first that these expressions simply have a number of separate meanings (a physical meaning, a moral meaning, etc.). However, research within formal semantics points toward a very different view. Such research suggests that all of these different uses can be explained by a single unified account of the meaning of modal expressions (Kratzer, 1977, 1981).

Difficult questions arise about the technical details of this account, but at its core lies a very simple idea. In any given context, people treat a certain set of possibilities as the relevant ones and regard the other possibilities as in some way irrelevant. Suppose we now refer to the set of relevant possibilities in any given context as the *domain*. We can then give a semantics for modal expressions in terms of this domain. For example, (2a) would mean something like (2b).

(2) a. It can be that *p*.
b. There is a possibility in the domain in which *p*.

Similarly, (3a) would mean something like (3b).

(3) a. It must be that *p*.
b. In all possibilities in the domain, *p*.

Analogous semantic clauses can be constructed for other modals.

On this view, there is no need to posit distinct meanings for the different uses of modal expressions. Instead, the differences arise simply because the domain is constructed differently in different contexts. In a context like the one found in (1a), the domain contains possibilities that don't violate physical laws. In (1b), it contains possibilities that don't violate moral requirements. In (1c), it contains possibilities in which you achieve your goals. Yet, though the domain is somewhat different in different contexts, the basic meaning of the modal expression itself remains constant.

Logic textbooks sometimes describe these different ways of constructing the domain just by giving a list of separate kinds of modals. This may leave the reader with the sense that any given modal has to fall neatly into one of these categories. In natural language, however, things tend not to be quite so tidy. Instead, we often find *impure modals* (Knobe & Szabó, 2013). That is, we find modals in which the domain is shaped by a number of different considerations: partly by physical law, partly by moral requirements, partly by goals, and so forth.

For example, in a modal like (4), it may seem at first that the domain is specifically shaped by the agent's goals.

(4) To get to Harlem, you have to take the A train.

However, it seems that moral considerations actually play a role as well. Thus, suppose that you could get to Harlem by getting on the G train, pulling out a gun, and then threatening to shoot someone unless the train went to Harlem. This option is so horribly immoral that it is seen as falling outside the domain. Thus, even if you could have achieved the goal in this way, sentence (4) will still be heard as true.

To sum up: Existing work in formal semantics has made important progress by positing a set of possibilities known as the *domain*. People tend to treat the possibilities that fall within this set as relevant and those that fall outside the set as irrelevant. The boundaries of this set are determined by a number of different considerations, but in general, there is a tendency whereby a possibility will be more likely to fall within the domain if it is morally good than if it is morally bad.

With this basic framework in place, we can now return to our original question. One approach would be to use the concept of a domain to spell out the intuitive idea that people regard certain possibilities as relevant and others as irrelevant. As we have seen, work in formal

semantics suggests that moral considerations play a role in which possibilities are included in the domain. If we now suppose that the domain plays a role in people's judgments about certain seemingly non-moral matters (freedom, causation, intentional action), we would then have an explanation of how moral considerations could end up impacting these judgments.

To really put that explanation to the test, we would have to spell out in detail precisely what role that domain plays in each of these judgments and then ask whether the resulting account can explain the patterns observed in existing studies. This task has been taken up within some existing research (Knobe & Szabo, 2013; Phillips & Cushman, 2017; Phillips & Knobe, 2018), but just for the moment, I want to put that whole issue to one side. The key point I want to emphasize is just that one approach to modeling the impact of moral judgment would be by using the framework initially introduced within research on modals. That is, one approach is to suppose that people are picking out a set of possibilities and then to ask how moral considerations play a role in determining whether or not a possibility falls within this set.

b. Probabilistic Sampling

Probabilistic sampling is a computationally tractable method for finding approximate answers to complex problems. It was originally developed within research in mathematics but is now an influential approach in computational cognitive science (Denison, Bonawitz, Gopnik & Griffiths, 2013; Vul, Goodman, Griffiths & Tenenbaum, 2014; for a review, see Icard, 2015)

To illustrate the basic idea, consider a problem you might encounter in your ordinary life. Tomorrow, there will be a party, and you want to make an educated guess about whether or not it will be fun. The problem is that you are not completely sure who will be there. Instead of having a list of people who will definitely attend, you just have a rough sense of how likely each person is to come. In a situation like this, what would be the best way of making a good guess about how fun the party will be?

At least in principle, one approach would be to consider every possible combination of people, estimate how much fun each of those combinations would be, and then weight each possible combination by the probability of it occurring. However, for problems that involve more than a very small number of variables, this strategy becomes completely unworkable. For example, if you know of ten different people who might or might not come, you would need to

consider more than a thousand different possibilities. There is no way that any actual human being would use this approach in deciding whether to go out with some friends for an evening.

An alternative strategy is therefore to use *probabilistic sampling*. Instead of exhaustively considering every single possibility, you would sample a certain number of possibilities and consider only those. Each time you took a sample, you would figure out how much fun it would be if that exact possibility arose. But here is the trick. You wouldn't just take samples arbitrarily; rather, you would sample possibilities in proportion to their probability. To guess how much fun the evening will be, you could then just take the average of the amount of fun in the different samples you considered. In other words, instead of considering more than a thousand different possibilities and weighting each one by its probability, you could arrive at an approximate answer just by thinking about a few specific possibilities and trying to guess how fun those specific possibilities would be. This is a far more plausible picture of the way human cognition actually works.

Thus far, we have been focusing on the idea that this method can be used in cases where a person is unsure what is going to happen and needs to consider a number of possible outcomes. However, precisely the same method can be used in other cases. In particular, we can use this method in cases where we already know what happened and we simply want to compare what actually happened to various alternative possibilities.

Suppose that we regard these other possibilities as relevant to different degrees. Some are highly relevant, some are entirely irrelevant, some have an intermediate status. We want our understanding of what actually happened to be influenced by these possibilities in proportion to their relevance (influenced more by the relevant possibilities, less by the irrelevant ones, etc.). We now face a problem that is more or less analogous in structure to the one we described above. We want to form an understanding of what happened that is informed by different alternative possibilities in proportion to their relevance. At least in principle, one could do this by considering every single possibility, taking into account its relevance, and then forming an overall weighted judgment. However, this sort of reasoning would not normally be feasible for actual human beings. Thus, we need to use a different approach. One obvious choice would be to turn to probabilistic sampling. Instead of considering every possibility and weighting it by its relevance, we simply sample from the possibilities.

This approach gives us a very different way of spelling out the idea that people regard morally good possibilities as more relevant. Specifically, we can spell out this idea in terms of the probability of being sampled. On this proposal, people have a higher probability of sampling a particular possibility when they regard it as morally good than when they regard it as morally bad.

To make this approach work, we need to introduce a somewhat novel view about the process of probabilistic sampling. One obvious initial assumption would be that people's sampling propensities reflect purely statistical representations (such as representations of the frequencies with which events occur in the world). The suggestion under discussion here is that we should abandon that assumption. Perhaps sampling propensities reflect not only statistical considerations but also *moral* considerations. As an example, consider again the case of believing that the electorate should respond with compassion. Suppose once again that you represent compassionate responses as not being very important from a purely statistical perspective (e.g., as having a low statistical frequency). The key hypothesis is that, even so, you might still have a high probability of sampling possibilities that involve compassionate responses, simply because you believe such responses to be morally right.

This framework then yields a new way of explaining the impact of moral consideration on people's judgments about apparently non-moral matters (freedom, causation, etc.). Perhaps people arrive at judgments about each of these matters through a process that involves sampling alternative possibilities. If moral consideration impact the probability that a given possibility is sampled, one would expect moral considerations to have an impact on the resulting judgments. Here again, the best way to evaluate this approach would be to try to spell out in real detail precisely how such a process would work. Within existing work, there have been attempts to do that for at least some kinds of judgments (Icard et al., 2017; Kominsky et al., 2015).

In the present context, however, the key point is just that this sampling-based explanation is actually very similar in form to the modality-based explanation we explored above. Of course, it might initially seem that the two explanations are radically different. One explanation uses set theory and draws on ideas from linguistic semantics; the other uses probability theory and draws on ideas from computational cognitive science. Yet this initial appearance is deceptive. The two frameworks can actually be seen as two ways of spelling out the same intuitive idea, namely, the

idea that people's moral judgments impact the degree to which they regard alternative possibilities as relevant.

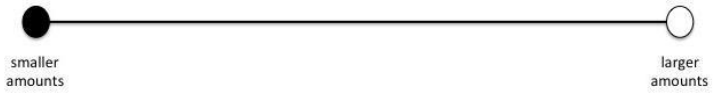
Thus, it would be a mistake just to have one stream of papers pursuing the first framework and another completely separate stream of papers pursuing the second. We need to think in a more serious way about how the two are related. We will be returning to that question shortly, but first, we need to put on the table yet another way of spelling out this judgment.

c. Normality

Research in a number of different areas has invoked the idea that people regard certain states of affairs as *normal* (Cialdini, Reno, & Kallgren, 1990; Dowty, 1979; Peysakhovich & Rand, 2015; Yalcin, 2016). In particular, it seems that people often make sense of the things that actually happen by comparing these things to what they regard as a normal state. In this way, people can determine whether the actual state of affairs departs in some way from the normal and, if so, in which direction.

Although the notion of normality can be deployed in a number of different ways, our focus here will be on cases in which people understand a state of affairs in terms of a point along a scale. For example, suppose that you are thinking about how much TV a particular person watches per day. You might understand this issue in terms of a scale that goes from very small amounts of TV up to very large amounts of TV. Now suppose that you have some rough intuition as to what counts as a normal amount of TV. You might then try to make sense of the specific person in question by comparing the amount she watches to the normal amount, seeing the amount she watches as 'normal,' 'less than normal' or 'more than normal.'

Existing research in this area has done a great deal to clarify people's ordinary understanding of scales (e.g., Kennedy & McNally, 2005). We now know a lot about how people think of different kinds of scales, how they map entities onto degrees on these scales, and how these degrees can be compared. Work in this area usually makes use of formal mathematical notation, but researchers also sometimes represent these scales visually using figures (e.g., Kennedy, 2007). Adopting that approach, we can represent the scale of amounts of TV as follows.

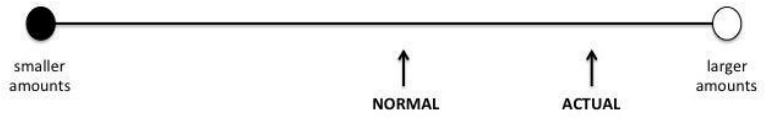


The black circle on the left signifies that the scale has a lower bound; the white circle on the right signifies that it has no upper bound. Within this broad framework, we can now pose a further question. How exactly do people determine which degree along the scale counts as the normal one?

It might at first seem that the answer is a matter of some purely statistical sort of judgment. Thus, one might think that the normal amount of TV is simply the average amount (or some other purely statistical measure along these same lines). However, this appears not to be the case. A number of different studies have explored people's ordinary judgments about normality, and all of them have arrived at the same conclusion. People's ordinary judgments about normality are impacted not only by statistical judgments but also by value judgments (Bear & Knobe, 2017; Wysocki 2018).

To illustrate, consider again our example of amounts of TV. People have a rough sense of how much TV the average person watches (a statistical judgment), and they also have a sense of the ideal amount of TV to watch (a value judgment). Strikingly, however, people's judgment about the normal amount of TV to watch is not simply equal to their judgment about the average. Rather, people pick out as the normal amount a point that is intermediate between the average and the ideal (Bear & Knobe, 2017). Thus, the average amount of TV is actually perceived not as normal but rather as abnormally large.

Consider now an agent who watches the average amount of TV. People will represent her as watching an amount that is greater than normal:



In this sense, people's value judgments impact their understanding of the amount of TV this agent actually watches. There might not be any impact of value judgments on their estimate of the absolute quantity the agent watches (expressed, e.g., as a number of hours), but there would still be an impact on judgments about whether this amount was normal, less than normal, or

greater than normal. To the extent that people think not in terms of absolute numbers but in terms of comparison to the normal, this could make all the difference.

Let us now turn to questions about the relationship between this framework and the two we discussed previously. One obvious view would be that there is no relationship at all. After all, the previous two frameworks were both concerned in some way with how people picked out the relevant possibilities. By contrast, the present framework is concerned with how people determine which degree along a scale is the normal one. It may seem, therefore, that we have simply switched over to a completely unrelated topic.

Admittedly, there is something right in this point. It is indeed correct that the topic we are taking up in the present section is quite different from the topics discussed in the previous ones, and almost all of the important discoveries from research on this topic would not be at all applicable to those. Still, it is hard to escape the sense that the specific claim we are making here is strikingly analogous to the one we made about probabilistic sampling.

To bring out the analogy as clearly as possible, we can articulate the two claims in a way that makes the parallel more evident. Here is one way of putting the claim about probabilistic sampling:

> When people are thinking about the actual situation, they often do so by considering various other possible situations. However, they do not treat all of these possible situations equally; they assign higher sampling propensity to some than to others. It might initially be thought that sampling propensity is simply proportional to some purely statistical property. However, that turns out not to be the case. Instead, sampling propensity is influenced both by statistical considerations and by moral considerations.

And here is one way of putting the point about normality:

> When people are thinking about an actual degree on a scale, they often do so by considering another degree along the same scale. However, they do not treat all degrees on the scale equally; they regard some degrees as more than normal than others. It might initially be thought that normality is simply proportional to some purely statistical property. However, that turns out not to be the case. Instead, perceived normality is influenced both by statistical considerations and by moral considerations.

Perhaps it would be possible to see these two points as just two special cases of a single more abstract principle. For example, something like this:

> When people are trying to understand something (a situation, a degree on a scale, etc.), they often do so by comparing it to an alternative (an alternative situation, an alternative degree on a scale). However, they do not treat all alternatives equally; they pick out certain particular alternatives as being especially worthy of consideration. These alternatives are picked out using both statistical and moral considerations.

Of course, I don't mean to suggest that this brief discussion constitutes any kind of solution to the problem. The aim is just to provide arguments for the view that there really is a problem here worth solving. In other words, my goal has been to show that these apparently unrelated frameworks are sufficiently similar that we face a real question as to how to understand the relation between them.

d. Relationship between the frameworks

Thus far, we have seen that the three formal frameworks are in some ways surprisingly similar and that, as a result, we face a question as to how to understand the relationship between them. To be honest, I do not know the answer to this question. Nonetheless, we can make at least some progress just by laying out a few plausible options.

1. One might think that these should be understood as three competing views about how to work out the details of a single basic vision. Thus, someone could say: 'All three of these views share a commitment to the basic idea that moral considerations impact their judgments about the relevance of different possibilities. Still, these views differ in their commitments about how exactly to implement that idea (e.g., in terms of sets vs. sampling propensities vs. scales). Future research should continue to ask whether the broader vision is on the right track but should also pit these different implementations against each other and ask which of them is most accurate.'

2. One might think that these views are not really in competition, in that different models could be preferable for understanding different psychological phenomena. For example, one might say: 'It is a basic fact about human cognition that moral considerations impact their judgments about the relevance of possibilities, but this basic fact manifests itself differently

when it comes to different phenomena. When people are using natural language, they use quantification over restricted domains, and this quantification is best understood in terms of sets. However, when people are not using natural language, the way they think is not best understood in terms of quantification over a restricted domain but rather in some other way (e.g., in terms of probabilistic sampling). Thus, future research should continue to explore this basic fact but should also examine the ways in which it works differently in different cases.'

3. One might think that these views are best understood as being fully compatible, in the sense that they aim to capture the same phenomenon at different levels of analysis. Hence, someone might say: 'It can be helpful for many purposes to think of people's cognitive processes as quantifying over a set of possibilities. However, this sort of analysis operates at a relatively high level, such that it doesn't say anything about the actual cognitive process people are using. (If we say that people are checking whether something holds of all the possibilities in a set, we presumably do not mean to imply that people go through every single one of these possibilities.) If we now want to descend to a lower level and think about the details of the actual algorithm people are using, we would need a different type of theory, and probabilistic sampling is one possible hypothesis about the workings of that lower level. Future research should therefore continue to explore these phenomena at both levels, understanding them at a more abstract computational level and also at a more detailed algorithmic level.'

We have been discussing three possible options, but I do not mean to suggest that these three are mutually exclusive and exhaustive. One can easily adopt a mix of these different options. For example, one might think that Option 2 describes the relationship between the modality-based account and the normality-based account, while at the same time thinking that Option 3 describes the relationship between the modality-based account and the probabilistic sampling account. Moreover, it is clear that these are not the only conceivable options. Further research may lead to the development of approaches that are not at all apparent at present.

In any case, to the extent that we are able to make real progress in answering this question, it will presumably not be by just contemplating these three options in the abstract. Rather, serious progress is likely to come only by trying to grapple with the application of these frameworks to the actual phenomena.

3. Back to the three effects

Within the existing literature, there has been a fair amount of attention to questions about whether these formal models can explain the three effects with which we began, but most of this work has had a somewhat different aim from the one we have been pursuing in the present chapter. Typically, a paper will pick out just one specific effect (e.g., the effect on causal judgments) and attempt to explain that effect using one specific framework (e.g., probabilistic sampling). The focus is then primarily on the details. That is, the paper aims to work out a particular implementation of the formal framework and show that this implementation can correctly capture the detailed pattern of people's judgments for the specific effect in question.

The present chapter takes up the opposite approach. Our emphasis will be on the big picture. Accordingly, we will be looking at the ways that these various different frameworks can help in explaining the various different effects. The aim is to see whether we can learn something from this more synoptic perspective that we would not have been able to learn by taking up just one framework or just one effect. Inevitably, this approach leads to a lack of engagement with the more detailed issues that have been the primary focus of existing research, but with any luck, we can make up for this loss of detail with a gain in a different sort of insight.

a. Freedom

Consider first people's judgments about freedom. It seems natural to approach this problem using the framework developed within research on modality. Indeed, claims about whether an agent acted freely seem to be quite closely related to claims that actually make use of natural language modals. Thus, a sentence of the form (1) seems closely related to a sentence of the form (2).

(1) The agent freely performed this action.

(2) The agent could have not performed this action.

Of course, difficult questions arise about precisely how to work out the details, but at some broad level, it seems that there is good reason to suspect that judgments about whether an agent performed an action freely have something to do with judgments about whether there are any relevant possibilities in which she did not perform the action.

This framework seems to adequately capture the most obvious features of our judgments about freedom. Suppose that an evil dictator tells me that he will cut off my hands unless I go to work on Saturday. If I then do go to work on Saturday, it seems that I do not do so freely. The framework can easily capture this judgment. Of course, one could conceive, at least in principle, of a possibility in which I simply choose not to work on Saturday and therefore lose my hands, but this possibility seems so outlandish as to be completely irrelevant. Thus, in all relevant possibilities, I do go to work on Saturday, and we therefore conclude that I could not have done otherwise and was acting unfreely.

We now arrive at a simple and elegant explanation for the impact of moral considerations observed in existing studies. Consider first the case in which the doctor obeys the chief of surgery's order, and if the doctor had disobeyed the order, the patient would have died. In this case, all possibilities in which the doctor disobeys are regarded as irrelevant, and the doctor's behavior is therefore classified as unfree. By contrast, consider the case in which the doctor obeys the chief of surgery and the patient dies, but if the doctor had disobeyed, the patient will have survived. In that case, the possibilities in which he disobeys are morally good. The moral properties of these possibilities lead us to regard them as relevant. For this reason, there are indeed relevant possibilities in which he does otherwise, and his action is classified as free.

This explanation has been proposed and developed by a number of different researchers and has found support in a variety of experimental studies (Knobe & Szabó, 2013; Kratzer, 2013; Phillips & Cushman, 2017; Phillips & Knobe, 2009; Young & Phillips, 2011). I am not aware of any claims that this framework fails to accurately predict or explain the data about people's freedom judgments in particular.

However, a question arises at the level of the bigger picture. The phenomena we observe in the people's judgments about freedom seem deeply related to phenomena we observe in judgments about other matters (causation, intentional action, etc.). Thus, it is not enough just to ask whether this framework correctly captures the phenomena in people's freedom judgments. We also need to ask whether it can capture the broader pattern of which this appears to be just one instance.

b. Causation

Considerable controversy remains about how to explain the impact of moral considerations on causal judgments. A variety of researchers have argued for some form of the possibility hypothesis (Blanchard & Schaffer, 2013; Halpern & Hitchcock, 2015; Icard et al., 2017; Kominsky et al., 2015), but a number of alternative hypotheses have also been proposed (Alicke, Rose & Bloom, 2011; Samland & Waldmann, 2016; Sytsma, Livengood & Rose, 2010). Existing work in this area has focused on looking in detail at the patterns in people's causal judgments to determine which of these hypotheses is actually correct (e.g., Livengood & Rose, 2016; Phillips et al., 2015).

I will not attempt here to review this existing work. Instead, I focus on a different question. Specifically, if we do opt for some form of the possibility hypothesis for the causation effect, what do we thereby learn about the big picture question as to how to understand these effects more broadly?

At the core of the possibility hypothesis for causal judgments is a very simple idea. People seem to make causal judgments by considering certain counterfactuals. Any process that impacts which counterfactuals people regard as relevant should therefore impact people's causal judgments. Thus, if moral considerations impact the degree to which people regard certain counterfactuals as relevant, we should expect an impact of moral considerations on causal judgments.

This basic approach can be applied quite straightforwardly to the experiment involving the professor and the pens (§1.2). Since the professor should not have taken a pen, the counterfactual in which she does not take a pen should be seen as especially relevant. By contrast, since the administrative assistant was in no way obligated to refrain from taking a pen, the counterfactual in which she does not take a pen should be seen as less relevant. If the perceived relevance of counterfactuals impacts causal judgment, one might then expect that the professor should be regarded as more causal than the administrative assistant.

A question now arises as to how to spell out this informal explanation in a more precise model. One approach would be to turn to again invoke the idea of a domain of possibilities. However, this approach immediately runs into a problem. In the case of freedom judgments, the key claim was that one of the conditions involved a possibility that was so far-fetched that it was not included in the domain at all. But that is not the structure of the case we are trying to understand here. Rather, in this case, neither possibility seems completely far-fetched or

irrelevant; it is just that one is even more relevant than the other. It is difficult to see how one could make sense of this sort of case in a framework in which we only have a dichotomous distinction between possibilities that fall inside vs. outside the domain.

Of course, in saying this, I don't at all mean to suggest that there is no hope for such an explanation. One might suggest that people's understanding of the administrative assistant's actions depends in part on their understanding of the professor's actions. Perhaps when the professor is not violating a norm, the possibility in which the administrative assistant behaves differently falls inside the domain, but when the professor is violating a norm, the possibility in which the administrative assistant behaves differently falls outside the domain. This suggestion is certainly a coherent one, and it could be investigated in further work.

However, it becomes much easier to make sense of this effect when we switch over to thinking in terms of sampling propensities. Then we no longer need to think in terms of a simple dichotomy. Instead, we can explain these phenomena straightforwardly in terms of differences in sampling propensity along a continuous scale. The possibility in which the administrative assistant behaves differently always has a moderate sampling propensity. Then the difference between conditions lies solely in the sampling propensity of the possibility in which the professor behaves differently. In one condition, this possibility also has a moderate sampling propensity; in the other, its sampling propensity is considerably higher.

The key point now is that research on causal judgments actually has implications for the broader question about how to understand the impact of moral judgment. There is strong reason to think that the causation effect is not completely unrelated to the freedom effect, and we therefore have reason to reject views that explain these two effects using two completely unrelated frameworks. One possibility would be to revise our understanding of the freedom effect; another would be to revise our understanding of the causation effect; a third would be to develop a more abstract account that allows us to see how these apparently different frameworks might actually be closely connected. In any case, it would be a mistake just to allow research in these two areas to continue on separately, with work on each area proceeding in isolation from the other.

c. Intentional action

The impact of moral considerations on intentional action judgments has generated considerably more controversy than either of the previous two we considered. Researchers have proposed a broad range of different hypotheses (Machery, 2008; Nichols & Ulatowski, 2007; Uttich & Lombrozo, 2010; see Cova, 2015, for a review of seventeen different hypotheses). Thus, the possibility hypothesis is just one of the many hypotheses that have been actively investigated for this effect.

Here too, the bulk of existing work aims to look in detail at the patterns of people's judgments to decide between these competing hypotheses. However, here again, I will not be reviewing that work. Instead, I ask how one would apply the possibility hypothesis in this case and how the application in this case might relate to the question of how to explain these effects more broadly.

The first thing to note is that attributions of intentional action are concerned not so much with an agent's behavior as with I that agent's *mental state*. Thus, the possibilities we will be exploring in this case will not be possibilities in which the agent performed a different behavior. Rather, they will be possibilities in which the agent had a different attitude toward the behavior.

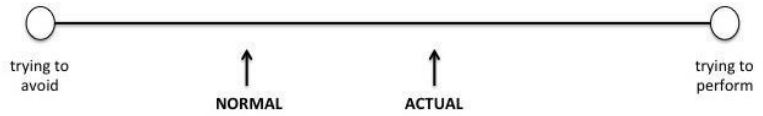
To begin with, we can use a continuous scale to represent the possible attitudes the agent could have had. On one end would be the attitude of an agent who is trying as hard as she can to avoid bringing about an outcome. On the other would be the agent who is trying as hard as she can to bring it about. Other attitudes can be represented as intermediate between these extremes.

We can now understand the concept of intentional action in part in terms of this scale. To the extent that an attitude falls toward the low end of the scale, people should be highly unlikely to regard it as intentional. To the extent that it falls toward the high end they should be highly likely to regard it as intentional. Then there is some vague threshold at which it crosses over from unintentional to intentional.

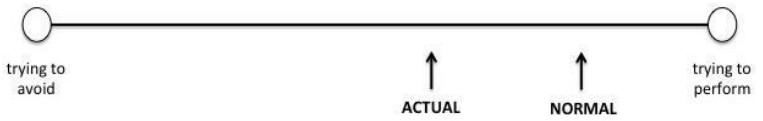
How exactly do people determine this threshold? The core idea is that it does not lie at some fixed and absolute point along the scale. Rather, the threshold is determined in part by the degree along the scale judged to be *normal* in any given case. The agent's actual attitude is

exactly the same in the harm case and the help case, but the normal attitude is different in the two cases, and as a result, people judge them differently.

In the harm case, the agent's actual attitude is at a higher point along the scale than the normal point. The agent is therefore regarded as strikingly willing to harm the environment. The agent's attitude falls above the threshold, and the action is classified as intentional.



By contrast, in the help case, the agent's actual attitude is at a lower point along the scale than the normal point. The agent is therefore regarded as strikingly reluctant to help the environment. The agent's attitude falls below the threshold, and the action is classified as unintentional.



We have been focusing here on one particular explanation for the intentional action effect. However, it should be noted that this has been a highly contested area of research, and numerous other explanations have been proposed (e.g., Machery, 2008; Nichols & Ulatowski, 2007; Uttich & Lombrozo, 2010). It is therefore possible that the explanation we have been discussing will turn out not to be correct. That said, a variety of recent experimental studies have provided evidence in favor of this explanation (Cova, Lantian, & Boudesseul, 2016; Phillips et al., 2015; Proft, Dieball, & Rakoczy, in press), so at the very least, there is strong reason to continue exploring it.

Let us now assume, if only for the sake of argument, that this explanation turns out to perfectly capture the patterns of people's intentional action judgments. What I want to suggest is that even then, we would still face a further question. After all, the effect we observe here is quite similar to the effects we observe for freedom and causation. Thus, even if we have a framework that does a good job of making sense of this one effect, we would have reason to be

dissatisfied unless that framework could also help us to see how this effect is connected with the other two.

Conclusion

We have been discussing three formal frameworks that have been proposed to explain the impact of moral considerations on people's apparently non-moral judgments. Although these frameworks might initially seem quite different from each other, I have argued that they are actually quite closely connected. In fact, I have suggested that they are best understood as three different ways of spelling out the same basic idea: the *possibility hypothesis*.

We have been focusing on the fact that this claim leaves us with some new theoretical puzzles. Within existing research, it is common to explain different effects of moral considerations using different formal frameworks. This research has led to many helpful advances, but I have been arguing that it also leaves us with a further, as yet unanswered question. Given that the formal frameworks are so closely connected, one wants to have some understanding of how to fit these various separate explanations into a unified account.

Yet there is also a more positive upshot of the claim defended here. The successes of each separate explanation provide some support for the specific formal framework in which it is formulated. However, if each of these frameworks is best understood as just one version of a single more general hypothesis, these successes also provide support for that general hypothesis. Thus, even if some or all of these more specific frameworks turn out to be mistaken, we now have fairly strong evidence that the possibility hypothesis itself is at least broadly on the right track.

We can therefore begin to explore more abstract questions that arise for the hypothesis more or less independently of the details of how is spelled out. We face questions about ultimate explanation (why would moral considerations affect representations of possibility in the first place?), about the role of these representations of possibility in our practical lives (how are our decisions affected when we regard certain possibilities as irrelevant?), and about the connection of these topics to broader issues in moral psychology. Research attention to these issues might first have emerged out of an attempt to understand certain relatively circumscribed questions involving judgments of freedom, causation and intentional action, but now that we have the

issues on the table, we can begin taking them up as topics worthy of investigation in their own right.

References

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. The Journal of Philosophy, 108(12), 670-696.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, 167, 25-37.

Blanchard, T., & Schaffer, J. (2013). Cause without default. Making a Difference, 1-29.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6), 1015.

Cova, F., Lantian, A., & Boudesseul, J. (2016). Can the Knobe Effect Be Explained Away? Methodological Controversies in the Study of the Relationship Between Intentionality and Morality. Personality and Social Psychology Bulletin, 0146167216656356.

Cova, F. (2015). The folk concept of intentional action: empirical approaches. *A Companion to Experimental Philosophy*.

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

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, 126(2), 285-300.

Dowty, D. (1979). *Word meaning and Montague grammar*. Dordrecht: Reidel.

Egré, P., & Cova, F. (2015). Moral asymmetries and the semantics of many. Semantics and Pragmatics, 8, 13-1.

Falkenstien, K. (2013). Explaining the effect of morality on intentionality of lucky actions: the role of underlying questions. Review of Philosophy and Psychology, 4(2), 293-308.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral psychology*, 2, 441-8.

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. The British Journal for the Philosophy of Science, 66(2), 413-457.

Icard, T. (2015). Subjective Probability as Sampling Propensity. Review of Philosophy and Psychology, 1-41.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80-93.

Kennedy, C. (1999). Projecting the adjective: The syntax and semantics of gradability and comparison. Routledge.

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

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 345-381.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63, 190-194.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psychology*, 2, 441-8.

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

Knobe, J., & Szabó, Z. G. (2013). Modals with a taste of the deontic. *Semantics and Pragmatics*, 6(1), 1-42.Chicago

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137, 196-209.

Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, 1(3), 337-355.

Kratzer, A. (1981). The notional category of modality. *Words, Worlds, and Contexts*, 38-74.

Kratzer, A. (2013). Modality for the 21st century. In *19th International Congress of Linguists* (pp. 181-201).

Livengood, J., & Rose, D. (2016). Experimental philosophy and causal attribution. *A Companion to Experimental Philosophy*, 434-449.

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

Ngo, L., Kelly, M., Coutlee, C. G., Carter, R. M., Sinnott-Armstrong, W., & Huettel, S. A. (2015). Two distinct moral mechanisms for ascribing and denying intentionality. *Scientific Reports*, 5.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language*, 22(4), 346-365.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, 24(5), 586-604.

Peysakhovich, A., & Rand, D. G. (2015). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3), 631–647.

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

Phillips, J. & Cushman, F. (2017) Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences.* 114, 4649-4654.

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

Phillips, J. & Knobe, J. (2018). The psychological representation of modality. *Mind & Language. 33*, 65-94.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, 145, 30-42.

Proft, M., Dieball, A. & Rakoczy, H. (in press). What is the cognitive basis of the side-effect effect? An experimental test of competing theories. *Mind & Language*.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, 156, 164-176.

Samland, J., Josephs, M., Waldmann, M. R., & Rakoczy, H. (2016). The role of prescriptive norms and knowledge in children's and adults' causal selection. *Journal of Experimental Psychology: General*, 145(2), 125.

Sripada, C. S., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind & Language*, 26(3), 353-380.

Sytsma, J., Livengood. J. & Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Science*, 43, 814-820.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116(1), 87-100.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4), 599-637.

Wysocki, T. (2017). Normality: A two-faced concept. Unpublished manuscript.

Yalcin, S. (2016). Modalities of normality. In N. Charlow & M. Chrisman (Eds.), *Deontic modals* (pp. 230–255). Oxford University Press.

Young, L., Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of Cognition and Culture*, 6(1-2), 265-278.

Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, 119(2), 166-178.