Folk Judgments of Causation[1]

Joshua Knobe
UNC-Chapel Hill

When scientists are trying to uncover the causes of a given outcome, they often make use of statistical information. Thus, if scientists wanted to know whether there was a causal relationship between attending philosophy lectures and learning philosophy, they might randomly assign students to either attend or not attend certain lectures and then check to see whether those who attended the lectures ended up learning more philosophy than those who did not.

A question now arises as to how ordinary people – people who have no formal training in the sciences – typically uncover the causes of particular outcomes. One popular answer to this question is that ordinary people use more or less the same techniques that scientists do (e.g., Gopnik et al. 2004; Kelly 1967; Woodward 2003). Of course, ordinary people do not actually write out equations or use precise statistical methods, but one might nonetheless claim that they uncover causal relations by looking in a more informal way at statistical information and that they actually look for pretty much the same sorts of statistical information that scientists do. At least within social psychology, this view is associated with the slogan *The person as scientist.*

Although this 'person as scientist' theory remains the dominant view both in philosophy and in psychology, recent years have seen the emergence of a new research program whose results point in a radically different direction. Research within this program has shown that people's causal judgments can sometimes be influenced by their *moral* judgments. In other words, when people are wondering about the causal relationships between events, their conclusions can be influenced by their beliefs as to whether those events are morally good or morally bad. At least on the surface, these results seem to serve as a challenge to the view that people assess causation by making something like a scientific judgment.

Yet researchers have not generally reacted by questioning prior assumptions about the nature of people's causal judgments. Instead, the usual view is that people truly are trying to make a purely statistical judgment but that certain processes are leading to 'distortions' or 'biases' in these judgments. The basic idea behind this view can be captured by the slogan *The person as bumbling scientist.* In essence, what it says is that people are engaged in an attempt to make scientific judgments but that they are messing up somehow and thereby allowing moral considerations to influence their intuitions.

My aim here is to offer an alternative hypothesis. I posit a single underlying mechanism that explains both the impact of statistical considerations and the impact of moral considerations. The claim, then, is that we should abandon the idea that causal

[1]I am grateful to Jonathan Weinberg and Hunt Stillwell for their suggestions regarding the basic idea at the root of this paper, to Christopher Hitchcock for numerous conversations regarding causal cognition, and to two anonymous referees for suggestions on an earlier version of this manuscript. Finally, I am grateful to Jim Woodward for going beyond the call of duty to provide extremely helpful in-depth comments on all aspects of the present paper.

judgments are fundamentally statistical and that the influence of moral considerations constitutes some sort of 'bias' or 'distortion.' In its place, we can adopt a theory according to which moral considerations truly do play a role in the fundamental mechanisms underlying causal judgments.

The role of statistical considerations

George writes out what he regards as a profound and original idea and sends it off to an academic journal… But there is a bad outcome. The paper gets a negative review and is therefore rejected. In a situation like this, one can easily imagine a person wondering what exactly caused the bad outcome. Was it something about the paper itself? Or something about the reviewer? Or a combination of the two?

Of course, there is a fairly obvious sense in which both facts about the reviewer and facts about the paper stand in some sort of causal relation to the problem that results, and one could well imagine that people might be content simply to trace out these various causal relations and understand how each of them works. Yet it appears that people do not actually proceed in that way. Instead, they seem to select certain particular factors and refer to those alone as 'causes,' while classifying all the others as mere 'background conditions' or 'enabling factors.' This process has come to be known as *causal selection*, and it plays an important role in folk judgments of causation.

A long tradition of research within social psychology has shown that the process of causal selection is sensitive in systematic ways to *statistical* considerations. So, for example, consider the following two cases:

> (Case 1) George sends his paper to a number of different journals and conferences, and they all reject it. Meanwhile, the reviewer accepts a number of other papers written by other authors.
>
> (Case 2) George sends his paper to a number of different journals and conferences, and they all accept it. Meanwhile, the reviewer rejects every single paper he is given.

Research shows that people's causal judgments about George's paper will depend on which of these two cases is the actual one (e.g., Hilton & Slugoski 1986; McArthur 1972). People will tend to say that the problem was caused by George's paper in Case 1, whereas they will tend to say that the problem was caused by the reviewer in Case 2.

It is really quite a striking fact that people respond in this way. After all, even if the reviewer was disposed to reject 99.99% of philosophy papers, one could still say that the bad outcome actually was caused by something about the paper – namely, the fact that it wasn't one of the .01% of papers that the reviewer would be inclined to accept. Yet the available research shows that people tend not to respond in that way. Instead, when faced with a case like this one, they say that the bad outcome *was* caused by something about the reviewer but *wasn't* caused by something about the paper. What we want to understand now is why exactly people take statistical considerations into account in this way.

The usual view within social psychology is that people's judgments in such cases are more or less analogous to the judgments one might make in the course of a scientific inquiry. Suppose that a scientist was trying to figure out whether the fates of academic papers were mostly due to something about the papers themselves or whether they were

mostly due to something about the individual reviewers. The first step would probably be to conduct an *analysis of variance* (ANOVA). One could give a lot of different papers to a lot of different reviewers and try to figure out what percentage of the total variance was explained by facts about the papers and what percentage was explained by facts about the reviewers. (As it happens, this experiment has actually been conducted. The answer is that a substantial percentage of the variance is explained by facts about the reviewers and almost none is explained by facts about the papers; Blackburn & Hakel 2006; Cole et al. 1981).

Of course, no one suggests that ordinary people make causal judgments by using precisely the same mathematical procedures that scientists use when calculating an ANOVA, but many researchers have suggested that we can think of the process underlying ordinary causal judgments as being *similar* in certain ways to the calculation of a full-fledged ANOVA (Försterling 1989; Kelley 1967).[2] To a first approximation, the claim is that people tend to attribute outcomes to whichever factor they think explains the greatest percentage of the variance. If they think that most of the variance is explained by facts about reviewers and almost none is explained by facts about the papers themselves, they will tend to say that the bad outcome was caused not by anything about the paper but solely by something about the reviewer.

The role of moral considerations

But it seems that things are not quite so simple. As a number of studies have shown, people's causal judgments can be influenced not only by statistical considerations but also by *moral* considerations (Alicke 1992; Cushman 2006; Knobe & Fraser 2008; Solan & Darley 2001). That is, when people are wondering whether *x* caused *y*, their judgments depend in part on whether they believe that *x* itself is morally good or morally bad.

Perhaps the best way of conveying the basic issues here is to give a simple example. In one recent experiment, subjects were given the following vignette:

> 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.

---

[2] Recent years have seen the emergence of more complex statistical frameworks (e.g., Pearl 2000; Spirtes et al. 2000), but unlike the ANOVA model, these frameworks do not purport to solve the problem of causal selection. It would be possible, then, to develop a model according to which judgments about the underlying causal structure were purely scientific but the process of causal selection then introduced an additional element that went beyond questions that could be solved using scientific methods alone (see, e.g., Hitchcock 2006).

> 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. (Knobe & Fraser 2008)

From a statistical perspective, the behavior of the faculty member and the behavior of the administrative assistant seem more or less the same. Both agents performed behaviors of quite ordinary sorts, and each behavior was related to the ultimate outcome in the same way. Yet subjects did not treat these two behaviors alike. They judged that the faculty member *did* cause the problem but that the administrative assistant *did not* cause the problem.

What we see here, apparently, is an impact of moral considerations on causal judgments. It seems that people classify the faculty member's behavior as *wrong* and that this classification then influences their judgments about whether his behavior caused the problem. The key question now is why exactly this effect arises.

One obvious way to react here would be to conclude that results like this one call into question the 'person as scientist' theory. Hence, one might say: 'If people really were behaving like scientists, they would not allow moral considerations to influence their judgments. So the available evidence now suggests that people actually aren't behaving like scientists but are instead engaged in some very different sort of inquiry.' But that has not been the usual reaction thus far. Instead, the usual reaction has been to suggest that there is a sense in which the original theory was actually right and it is the experimental subjects themselves who are wrong. That is to say, researchers react to these results by suggesting that subjects truly are trying to engage in a purely statistical inquiry but that some further process is *interfering* here and allowing moral judgments to shape people's responses.

One way to address this issue would be to look in detail at a number of specific hypotheses about precisely how people's moral judgments could interfere with the proper operation of their causal reasoning. Some researchers have suggested that the effects here are due to a motivational bias (Alicke 1992, forthcoming); others have suggested that the effects are due to pragmatics (Driver 2008a). We could examine each of these hypotheses in turn, looking for confirming or disconfirming evidence. This strikes me as an extremely valuable project (and one I have attempted elsewhere; Hitchcock & Knobe 2008.; Knobe & Fraser 2008), but I will not be pursuing it here.

Instead, I want to proceed by offering a positive hypothesis about the underlying psychological mechanisms that generate these effects. On this hypothesis, there is no sense in which statistical considerations truly do play a role in the fundamental competence while moral considerations are merely a 'distortion' or 'bias.' On the contrary, the hypothesis suggests that *the very same process* explains the use of both statistical and moral considerations.

The classification of counterfactuals

In particular, I want to suggest that both of these effects can be explained in terms of a quite general theory about people's capacity for counterfactual reasoning. Hence, the suggestion will be that statistical and moral considerations end up having an impact

on people's causal judgments because they have an impact on the way people reason about counterfactuals in general.

The first thing to note here is that our capacity for counterfactual reasoning not only allows us to distinguish between counterfactuals that are true and those that are false but also allows us to distinguish between counterfactuals that are worth considering and those that we would do better simply to ignore. Thus, suppose that I end up getting a bad grade on an important test. I might immediately begin considering certain counterfactuals: 'What if I had studied harder?' 'What if I had chosen not to go to that party last night?' And many other similar thoughts might fill my mind. But there are also counterfactuals that I would regard as silly and not worth considering. I would never think: 'What if the teacher had been struck by a meteor before she finished grading my exam?'

It is an important fact about human cognition that we are able to focus in this way on the relevant counterfactuals and ignore the irrelevant ones. Counterfactual thinking can help us to solve many practical problems, but if we ended up reflecting in detail on any counterfactual that came to mind, we would never be able to get anything done. Ideally, then, our minds would somehow manage to focus in on the relevant counterfactuals and suppress all further thoughts about the irrelevant ones. It seems that human beings actually are surprisingly good at accomplishing this task.

There has been a great deal of experimental research in psychology on people's capacity for counterfactual reasoning, and a considerable amount is now known about the precise conditions under which people do and do not consider various counterfactuals. Here, however, we can proceed by borrowing just three basic principles from this vast literature:

- The first principle says that people are inclined to consider counterfactuals in which events of statistically *unusual* types are replaced by events of statistically *usual* types (Kahneman & Tversky 1982; Roese 1997). Example: if a student hands in her paper on a roll of toilet paper, we will be inclined to think about what would have happened if she had handed it in on computer paper instead.
- The second principle says that people are inclined to consider counterfactuals in which *bad* events are replaced by *good* events (Kahneman & Miller 1986; Read 1985). Example: if a committee makes a bad decision, we will be inclined to think about what would have happened if it had instead made a good decision.
- The third principle establishes a kind of default. It says that, unless there is some specific reason to think about a given counterfactual, people will be inclined to classify it as irrelevant and not give it any further thought.[3]

In articulating this third principle, I depart in certain respects from the traditional way of thinking about these phenomena within psychology. The traditional view is that certain types of counterfactuals just never occur to people in the first place. The hypothesis

[3] Ultimately, it might be possible to unify these three principles in the single claim that people regard a counterfactual as relevant to the extent that it replaces *abnormal* things with *normal* ones (Hitchcock & Knobe 2008). The key assumption would be that people have a notion of 'normality' that includes both statistical and moral elements.

being offered here is that there is actually something more complex going on. It is not just that certain counterfactuals never occur to people; it is that these counterfactuals are actually *classified as irrelevant.* To see the force of this claim, consider what might happen if we actively intervened in in a student's life and forced him to consider what would have happened if his teacher had been struck by a meteor. (One way to arrange such an intervention would be to ask the student to write a detailed essay on the topic.) After our intervention was complete, the student would have very vivid and definite views about precisely what would have occurred under the specified counterfactual conditions. Nonetheless, the hypothesis is that the student would classify the whole counterfactual as irrelevant and that his beliefs about it would therefore have little impact on any further cognitive processes he might undergo. Hence, if we later asked him to think about how he might avoid getting bad grades in the future, he would not begin by thinking: 'Well, I would have avoided this bad grade if my teacher had only been struck by a meteor…' And similarly for any other aspect of cognition. No matter where we look, we will find a certain resistance to considering counterfactuals that have been classified as irrelevant.

Ultimately, the hope is that we can draw on these very general claims about people's use of counterfactuals to explain the puzzling experimental results concerning people's judgments of causation. It is important to emphasize, however, that our claims about people's use of counterfactuals do not themselves depend on evidence from studies of causal reasoning. Instead, these claims are based on *independent* evidence. (The experimental studies look directly at people's counterfactual reasoning, not at their judgments of causation.) Our overall theory can therefore draw on two different sources of experimental evidence. First we look at studies of counterfactual reasoning and thereby construct a general theory about how people use counterfactuals; then we take this general theory and try to use it to explain the results of experiments about people's causal judgments.

Explaining the effects

To get from this general theory of counterfactuals to specific predictions about causal judgments, we need to introduce an additional assumption, namely, the assumption that people arrive at causal judgments by making use of counterfactuals.

Fortunately, we have good reason to believe that this assumption is correct. A wide variety of theories of causal judgment suggest that these judgments actually do rely on counterfactuals in one way or another (e.g., Collins et al. 2001; Hitchcock 2008; Lewis 2000; Woodward 2003). These theories differ from each other in a number of important respects, but those differences will not be relevant here. Instead, we will simply be relying on the basic claim that people make judgments about whether a given event caused an outcome by considering counterfactuals in which that event does not occur.

Armed with this assumption, we can now reexamine the experimental results concerning people's causal judgments. The aim will be to show that it is possible to explain the patterns we observe in these judgments by drawing on a general theory of counterfactual reasoning.

First, consider the role of statistical considerations. Our example here was the paper that had been submitted for review at an academic journal. The paper, let us

suppose, is good enough that it would normally be accepted, but it has been sent to a reviewer who has a tendency to reject almost every manuscript he receives. The question now is how people will determine what caused the bad outcome.

The thing to focus on here is the general principle that people tend to consider counterfactuals in which events of *unusual* types are replaced by events of *usual* types. Since the reviewer is taking a very unusual approach to the manuscript, people immediately consider the counterfactual in which he takes a more usual approach. That is, they consider a counterfactual of the form:

(1a) If the reviewer had applied a more ordinary standard to the manuscript…

The evaluation of this counterfactual then leads (in accordance with whichever theory turns out to be correct) to a judgment that the bad outcome was caused by the reviewer's unusual standard.

But there is also another aspect to the situation. We have been assuming that the reviewer does not reject absolutely all manuscripts and that there was therefore some way of writing the paper that would have led it to be accepted. (For concreteness, we might suppose that the paper would have been accepted if it had offered fulsome praise for the reviewer's own prior work.) Suppose, then, that people began wondering whether the bad outcome was actually caused by something about the paper – namely, the fact that it wasn't one of the .01% of papers that the reviewer would have accepted. To address this question, they would have to consider the counterfactual:

(1b) If the paper had been one of the .01% that fulsomely praised the reviewer's prior work…

But there is no principle that picks out this counterfactual as a relevant one. Hence, the counterfactual is classified as irrelevant, people do not give it any further consideration, and the properties of the paper itself don't end up being regarded as causes of the outcome.

A similar approach can be applied to understanding the role of moral considerations. Our example here involved a professor and an administrative assistant who each take a pen from the receptionist's desk. By the second principle laid out above, people should immediately be drawn to counterfactuals that involve changing bad events to good ones. Thus, they should be drawn to consider counterfactuals of the form:

(2a) If Professor Smith had not taken a pen…

And they should thereby end up concluding that Professor Smith's decision to take a pen was a cause of the outcome.

But now suppose people begin wondering whether the outcome was also caused by the administrative assistant's behavior of taking a pen. They would have to consider counterfactuals of the form:

(2b) If the administrative assistant had not taken a pen…

But there is no principle that would lead people to classify this second counterfactual as relevant. It is therefore classified as irrelevant and blocked from playing further roles in cognition. Ultimately, people do not end up concluding that the administrative assistant's behavior caused the outcome.

What we have here is a rough sketch of an explanation of the effects described above. Clearly, more work will be needed before this explanation can be considered complete, but it should be possible to see, at least in outline, how the various experimental results are to be explained. Above all, it should be clear that the explanation being offered here departs quite radically from the 'person as scientist' theory. If this explanation is on the right track, people's judgments may sometimes mimic the results of a systematic ANOVA, but the basic logic underlying their responses is fundamentally different from anything one might find in a purely statistical analysis. In particular, it seems that the very same process that allows people's judgments to be affected by statistical information also allows them to be affected by moral considerations.

Conclusion

The explanation being offered here has a somewhat unusual character, and it may therefore be helpful to say a few additional words about how it is supposed to work and how it contrasts with other explanations that have been offered for the same phenomena.

In thinking about patterns in folk judgments, researchers are often drawn to a mode of thought that might be called *teleological*. That is, researchers are often drawn to the thought that folk judgments must be serving some sort of purpose in people's lives and that we can gain an understanding of why people make these judgments in the way they do by thinking about how they thereby serve that purpose. This mode of thought is especially tempting in cases, like the one under discussion here, in which people's judgments show highly complex patterns. There is an almost overwhelming tendency to suppose that all of this complexity must have arisen because it helps people to accomplish some important purpose.

It seems clear that, this sort of thinking is at work in the 'person as scientist' theory. The basic intuition there is that the point of making causal judgments is to achieve a kind of proto-scientific understanding of the world. If the only considerations relevant to that sort of understanding are the statistical ones, then it is assumed that the underlying competence will only take statistical considerations into account. Any use of other sorts of considerations must involve some sort of interference with the proper workings of the mechanism.

On the view presented here, by contrast, it is somewhat difficult to see precisely what purpose the underlying competence might be serving. Hence, a person might ask: 'Why on earth would someone mix together statistical and moral considerations in this complex way? What possible purpose could all of this processing really serve?' If no answer was forthcoming, such a person might conclude that moral considerations must not be playing a role in the competence after all.

My response to this worry is to reject the whole idea that people's underlying competence should be understood as the optimal way of achieving some particular purpose. After all, it is not as though this competence was designed by an engineer who started from scratch and simply tried to create a mechanism that could do the best possible job of generating causal judgments. On the contrary, the competence is best understood as something cobbled together from parts that originally served a different purpose. (Think of the way people sometimes light a fire by using newspaper as

kindling. The newspaper is covered in writing – but not because that writing in any way contributes to the function of lighting fires.)

When we consider the matter from this latter standpoint, it is not at all difficult to see why statistical and moral considerations play the role they do. It is not that these considerations came to play a certain role because they could thereby contribute to the purpose of people's causal judgments. Rather, the use of these considerations is simply built into the fundamental mechanisms that subserve people's counterfactual reasoning. Any aspect of human cognition that makes use of counterfactuals will be affected in some way by the structure of these mechanisms. Since causal judgments make use of counterfactuals, and since moral considerations play a role in the mechanisms underlying counterfactual reasoning, moral considerations end up playing a role in causal judgments as well.

References

Alicke, M. (1992), Culpable causation, *Journal of Personality and Social Psychology*, 63, 368-378.

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

Blackburn, J. & Hakel, M. (2006), An examination of sources of peer-review bias, *Psychological Science*, 17, 378-382.

Cole, S., Cole, J., & Simon, G. (1981), Chance and consensus in peer review, *Science*, 214, 881-886.

Collins, J., Hall, N., and Paul, L. A. (2001), *Causation and counterfactuals*. Cambridge, Mass.: MIT Press.

Cushman, F. (2006), Judgments of morality, causation and intention: Assessing the Connections, Unpublished manuscript. Harvard University.

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

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

Försterling, F. (1989), Models of covariation and attriution: How do they relate to the analogy of analysis of variance?, *Journal of Personality and Social Psychology*, 57, 615-625.

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

Hilton, D. & Slugoski, B. (1986), Knowledge-based causal attribution: The abnormal conditions focus model, *Psychological Review*, 93, 75-88.

Hitchcock, C. (2008), Token causation, Unpublished manuscript. California Institute of Technology.

Hitchcock, C. (2006). Three concepts of causation. *Philosophy Compass* 2.

Hitchcock, C. & Knobe, J. (2008), Cause and norm, Unpublished manuscript. California Institute of Technology.

Kahneman, D. & Miller, D. (1986), Norm theory: Comparing reality to its alternatives, *Psychological Review*, 80, 136-153.

Kahneman, D. & Tversky, A. (1982), The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky. (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201-210), Cambridge, UK: Cambridge University Press.

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

Lewis, D. (2000), Causation as influence, *Journal of Philosophy*, 97, 182-198.

McArthur, L. (1972), The how and what of why: Some determinants and consequences of causal attribution, *Journal of Personality and Social Psychology*, 22, 171-193.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. New York: Cambridge University Press.

Read, D. (1985), Determinants of relative mutability, Unpublished research, University of British Columbia, Vancouver, Canada.

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

Solan, L. & Darley, J. (2001), Causation, contribution, and legal liability: An empirical study, *Law and Contemporary Problems*, 64, 265-298.

Spirtes, P., Glymour, C., and Scheines, R. 2000. *Causation, prediction and search, 2nd edn.* New York: MIT Press.

Woodward, J. (2003), *Making things happen*. Oxford: Oxford University Press.