Forthcoming in Walter Sinnott-Armstrong (ed.), *Moral Psychology*

Causal Judgment and Moral Judgment: Two Experiments[1]

Joshua Knobe
University of North Carolina─ Chapel Hill

Ben Fraser
Australian National University, Research School of the Social Sciences

Note: This paper is forthcoming in a volume on moral psychology, where it will be included as a response to a chapter by Julia Driver. However, the paper is written in such a way that it should be easily accessible to anyone interested in the relationship between causal judgment and moral judgment ─ including those who are not familiar with Driver's work.

It has long been known that people's causal judgments can have an impact on their moral judgments. To take a simple example, if people conclude that a behavior caused the death of ten innocent children, they will therefore be inclined to regard the behavior itself as morally wrong. So far, none of this should come as any surprise.

But recent experimental work points to the existence of a second, and more surprising, aspect of the relationship between causal judgment and moral judgment. It appears that the relationship can sometimes go *in the opposite direction*. That is, it appears that our moral judgments can sometimes impact our causal judgments. (Hence, we might first determine that a behavior is morally wrong and then, on that basis, arrive at the conclusion that it was the cause of various outcomes.)

There is still a certain amount of debate about how these results should be interpreted. Some researchers argue that the surprising results obtained in recent studies are showing us something important about people's concept of causation; others suggest that all of the results can be understood in terms of straightforward performance errors.

Driver provides an excellent summary of the existing literature on this issue,[2] but she also offers a number of alternative hypotheses that threaten to dissolve the debate

[1] We are grateful to Christopher Hitchcock for many hours of valuable conversation on these issues.
[2] Driver's summary focuses especially on the competing theories of Alicke (1992) and Knobe (2005). For further experimental evidence, see Alicke (2000), Alicke et al. (2004), Knobe (forthcoming) and Solan and Darley (2001). For philosophical discussions, see Beebe (2004), Gert (1988), Hart and Honore (1985),

entirely. These hypotheses offer ways of explaining all of the experimental data without supposing that moral judgments play any real role in the process that generates causal judgments. If her alternative hypotheses turn out to be correct, we will be left with no real reason to suppose that moral judgments can have an impact on causal judgments.

We think that Driver's hypotheses are both cogent and plausible. The only way to know whether they are actually correct is to subject them to systematic experimental tests. That is precisely the approach we have adopted here.

The Problem

Driver introduces the basic problem by discussing a few cases from the existing literature. Here is one of the cases she discusses:

> Lauren and Jane work for the same company. They each need to use a computer for work sometimes.
>
> Unfortunately, the computer isn't very powerful. If two people are logged on at the same time, it usually crashes.
>
> So the company decided to institute an official policy. It declared that Lauren would be the only one permitted to use the computer in the mornings and that Jane would be the only one permitted to use the computer in the afternoons.
>
> As expected, Lauren logged on the computer the next day at 9:00 am.
>
> But Jane decided to disobey the official policy. She also logged on at 9:00 am.
>
> The computer crashed immediately. (Knobe 2005; discussed in Driver this volume)

In this case, people seem more inclined to say that Jane caused the computer crash than they are to say that Lauren caused the computer crash. Yet Jane's behavior resembles

---

Hitchcock (2005), Knobe and Fraser (forthcoming), McGrath (forthcoming), Thomson (2003), Woodward (2003).

Lauren's in almost every way. The key difference between them is a purely normative one: Jane violated one of her obligations, whereas Lauren did not. Thus, it appears that people's normative judgments may be having some influence on their causal judgments.[3]

Driver's question is about how to understand what is going on in cases like this one. Do people's judgments about obligations, rights, etc. actually serve as input to their judgments about causal relations?

Morality and Atypicality

Driver's first suggestion is that it might be possible to explain all of the puzzling results by appealing to the concept of *atypicality*. Some behaviors are fairly common or ordinary, others are more atypical. Perhaps we can explain the results of existing experiments if we simply assume that people have a general tendency to pick out atypical behaviors and classify them as causes.

With this thought in mind, we can revisit the story of Jane and Lauren. We noted above that Jane's behavior differs from Lauren's in its moral status, but it seems that we can also identify a second difference between the two behaviors. Jane's behavior seems quite atypical for a person in her position, whereas Lauren's behavior seems perfectly common and ordinary. So perhaps people's tendency to pick out Jane's behavior and classify it as a cause has nothing to do with its distinctive moral status. It might be that people simply classify Jane's behavior as a cause because they regard it as atypical.

This point is well-taken. The immoral behaviors in existing experiments were always atypical. Thus, Driver's hypothesis explains all of the existing results just as well as the hypothesis that moral judgments really do have an impact on causal judgments. To decide between the competing hypotheses, we will therefore need to conduct an additional experiment.

What we need now is a case in which two behaviors are equally typical but one is morally worse than the other. Here is one such case:

---

[3] This point has occasionally been made with regard to causation by omission (Beebe 2003; McGrath forthcoming; Thomson 2003; Woodward 2004). But the effect does not appear to have anything to do with omissions specifically. Normative judgments appear to affect causal judgments even in cases like this one where we are not concerned with omissions in any way.

> 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.
>
> 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.

In this case, the professor's action and the administrative assistant's action are both typical, but only the professor's action is in any way reprehensible. What we want to know is whether this small difference in perceived moral status can ─ all by itself, with no help from typicality judgments ─ have any impact on people's causal judgments.

To address this question, we ran a simple experiment. All subjects were given the story of the professor and the administrative assistant. They were then asked to indicate whether they agreed or disagreed with the following two statements:

- 'Professor Smith caused the problem.'
- 'The administrative assistant caused the problem.'

The results showed a dramatic difference. People agreed with the statement that Professor Smith caused the problem but disagreed with the statement that the administrative assistant caused the problem.[4] Yet the two behaviors seem not to differ in their typicality; the principal difference lies in their differing moral statuses. The results therefore suggest

[4] Subjects were 18 students in an introductory philosophy class at University of North Carolina-Chapel Hill. The order of questions was counterbalanced, but there were no significant order effects. Each subject rated both statements on a scale from -3 ('not at all') to +3 ('fully'), with the 0 point marked 'somewhat.' The mean rating for the statement that the professor caused the problem was 2.2; the mean for the statement that the assistant caused the problem was -1.2. This difference is statistically significant, $t(17) = 5.5$, $p < .001$.

that moral judgments actually do play a direct role in the process by which causal judgments are generated.

Conversational Pragmatics

When researchers want to understand people's concept of causation, the usual approach is to look at how people apply the English word 'cause.' But it should be clear that this method is a fallible one. There are certainly cases in which people's use of words in conversation can diverge from their application of concepts in private thought. After all, the point of conversation is not simply to utter sentences that correspond to one's beliefs. People are also concerned in an essential way with the effort to provide information that is relevant and helpful to their audience members. Thus, if people think that using the word 'cause' in a given case might end up giving audience members the wrong impression, they might refuse to use that word for reasons that have nothing to do with a reluctance to apply the corresponding concept in their own private thoughts. Here we enter the domain of *conversational pragmatics*.[5]

Driver's second major suggestion is that it might be possible to account for the apparent role of moral considerations in causal judgments simply by appealing to the pragmatics of conversation. The basic idea here is simple and quite plausible. When we offer causal explanations, our aim is not simply to utter sentences that express true propositions; we are also engaged in an effort to say things that prove relevant and helpful to our audience. It seems clear, moreover, that moral considerations will often play an important role in this pragmatic aspect of conversation. When a person's house has just been burned down, it won't do to give him just any cause of the fire. He will want a causal explanation that helps him to figure out who is morally responsible for what happened. In this way, moral considerations can affect the *speech act* of offering a causal explanation even if they play no role in people's underlying *concept* of causation.

We certainly agree that conversational pragmatics can have an impact on people's use of causal language. The only question is whether the apparent connection between

[5] Note that our concern in this section is only with the pragmatics of *conversation*. Even if the effect under discussion here has nothing to do with conversational pragmatics, one might still argue that it is 'pragmatic' in a broader sense, i.e., that it arose because it enables us to achieve certain practical purposes.

moral judgments and causal judgments is due to pragmatics alone. In other words, the question is whether this connection is due *entirely* to pragmatics (so that it would simply disappear if we eliminated the relevant pragmatic pressures) or whether the connection is also due in part to *other processes* (so that it would persist even if we could somehow eliminate the pragmatics entirely).

One way to investigate this issue is to construct a case in which conversational pragmatics alone gives us no reason to specifically pick out the morally bad behavior and classify it as a cause. Here is one such case:

> Claire's parents bought her an old computer. Claire uses it for schoolwork, but her brother Daniel sometimes logs on to play games. Claire has told Daniel, "Please don't log on to my computer. If we are both logged on at the same time, it will crash".
>
> One day, Claire and Daniel logged on to the computer at the same time. The computer crashed.
>
> Later that day, Claire's mother is talking with the computer repairman. The repairman says, "I see that Daniel was logged on, but this computer will only crash if two people are logged on at the same time. So, I still don't see quite why the computer crashed."

The morally bad behavior in this case is Daniel's act of logging on, but the conversational context is constructed in such a way that there are no pragmatic pressures to perform the speech act of asserting that this morally bad behavior was the cause. In fact, all of the pragmatic pressures go in the opposite direction. Even though Daniel's act of logging on is a morally bad behavior, it would be inappropriate in this conversation to mention it in a causal explanation. The key question now is how a speaker would react in such a case.

If the entire connection between moral judgments and causal judgments were due to conversational pragmatics, the connection should simply disappear in cases like this one. That is, the speaker should be left with no inclination at all to specifically pick out the immoral behavior and classify it as a cause.

But there is another possible way of understanding what is going on here. Perhaps the connection between moral judgments and causal judgments is not merely a matter of pragmatics but actually reflects something fundamental about the way people ordinarily think about causation. On this hypothesis, some connection should persist even in cases like the one under discussion here. Even when it seems pragmatically inappropriate to *say* that the immoral behavior was the principal cause, people should still *think* that the immoral behavior was the principal cause. Thus, the mother should think: 'Daniel's act of logging on truly was the cause of the computer crash; it just wouldn't be appropriate to mention that in this conversation.'

To decide between these hypotheses, we ran a second experiment. All subjects were given the story of Claire and Daniel. Each subject was then asked two questions. One question was about what explanation would be most *appropriate in the conversation*; the other was about what the mother *actually believes*.[6]

The results were simple and striking. The vast majority of subjects (85%) responded that it would be most appropriate in the conversation to explain the crash by saying that Claire was logged on.[7] But intuitions about what the mother actually believed did not correspond to intuitions about what it would be most appropriate to say. Instead, subjects responded that the mother would believe that Claire *did not* cause the crash but that Daniel *did* cause the crash.[8]

Ultimately, then, it seems that there is more going on here than can be explained by pragmatics alone. Pragmatics can determine what would be appropriate to say in a given conversation. But even in conversations where it would clearly be inappropriate to mention the morally bad behavior, people insist that an observer would specifically pick out that behavior and regard it as a cause. This result suggests that moral considerations

---

[6] Subjects were first asked to say which of two replies would be more appropriate in the conversation. The two options given were 'Daniel was logged on' and 'Claire was logged on.' Subjects were then asked to rate the degree to which the mother actually believes each of two sentences. The two sentences were 'Daniel caused the computer crash' and 'Claire caused the computer crash.'

[7] This percentage is significantly greater than what would be expected by chance alone, χ^2 (1,47) = 24.1, $p < .001$.

[8] Subjects rated each statement on a scale from -3 ('not at all') to +3 ('fully'), with the 0 point marked 'somewhat.' The order of questions was counterbalanced, but there were no significant order effects. The mean rating for the statement that Claire caused the computer crash was -1.3; the mean for the statement that Daniel caused the computer crash was 1.6. This difference is statistically significant, $t(40) = 6.2$, $p < .001$.

are not merely relevant to the pragmatics of conversation but actually play a fundamental role in the way people think about causation.

References

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

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

Alicke, M.D., Davis, T.L., & Pezzo, M.V. (1994). A posteriori adjustment of a priori decision criteria. *Social Cognition*, 12, 281-308.

Beebee, H. (2004). Causing and nothingness. In L.A. Paul, E.J. Hall and J. Collins, eds., *Causation and Counterfactuals.* Cambridge , Mass.: MIT Press), 291-308

Gert, B. (1988). *Morality: A new justification of the moral rules*. New York: Oxford University Press.

Hart, H.L.A., and Honoré, T. (1985). *Causation in the Law*, 2nd ed. Oxford: Clarendon.

Hitchcock, C. (2005). Token causation. Unpublished manuscript. California Institute of Technology.

Knobe, J. (forthcoming). Cognitive processes shaped by the impulse to blame. *Brooklyn Law Review*.

Knobe, J. (2005). Attribution and normativity: A problem in the philosophy of social psychology. Unpublished manuscript. University of North Carolina-Chapel Hill.

Knobe, J. & Doris, J. (forthcoming). Strawsonian variations: Folk morality and the search for a unified theory. *Rethinking Moral Psychology.*

McGrath, S. (forthcoming). Causation by omission. *Philosophical Studies*.

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

Thomson, J. (2003). Causation: Omissions, *Philosophy and Phenomenological Research*, *66*, 81.

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