Brooklyn Law Review

Volume 71 | Issue 2 Article 5

2006

DIALOGUE: Cognitive Processes Shaped by the Impulse to Blame

Joshuea Knobe

DIALOGUE

Cognitive Processes Shaped by the Impulse to Blame

Joshua Knobe[†]

In his incisive and thought-provoking paper "Cognitive Foundations of the Impulse to Blame," Lawrence Solan points to a surprising fact about the cognitive processes underlying attributions of blame.[1] This surprising fact is that almost all of the processes that we use when trying to determine whether or not a person is blameworthy are also ones that we sometimes use even when we are not even considering the issue of blame.[2] Only a very small amount of processing is used *exclusively* when we are interested in questions of blame.

This point can be made vivid with a simple example. Suppose that we witness a terrible accident and then assign an investigator to answer the question: "Why did this accident occur?" This investigator spends many months gathering evidence, formulating hypotheses, and considering arguments of various types. Finally, he comes back with a definite answer. And now suppose we tell him that we also want an answer to a second question, namely: "Was anyone to blame for this accident?" The investigator probably won't have to spend another few months answering this new question. It appears that almost all of the work has already been done; the investigator can simply take the results he has already obtained, do a little extra thinking, and come up with an answer.

[†] Princeton University. I am grateful to Lawrence Solan and Gilbert Harman for helpful comments on an earlier draft.

[1] Lawrence M. Solan, *Cognitive Foundations of the Impulse of Blame,* 68 BROOK. L. REV. 1003 (2003).

[2] *Id.* at 1004.

Solan provides support for this initial intuition through a sophisticated analysis of the cognitive processes that underlie attributions of blame. Specifically, he shows that attributions of blame rely in a crucial way on judgments about *mental states* and about *causal relations*.[3] He then shows that we would have made these very judgments anyway, even if we had not been concerned with questions of blame.

Solan also offers a tentative hypothesis about why the cognitive processes that underlie attributions of blame overlap in this way with the cognitive processes used in other contexts. He suggests that perhaps human beings first began using these processes for some entirely separate purpose – e.g., because they served a useful role in predicting and explaining behavior – and that these processes then came to be used in blame attributions as well.[4]

Solan is calling our attention to a very important phenomenon here, but I want to suggest that we ought to draw almost exactly the opposite conclusion about it from the one he has drawn. The phenomenon is that nearly all of the cognitive processes that we use when assessing blame are also processes that we use when the question of blame does not even arise. Solan's conclusion is that blame has had a relatively small impact on the capacities that underlie our cognitive processes.[5] I would draw the opposite conclusion: blame has had such a pervasive influence on our cognitive capacities that, even when we are not specifically interested in questions of blame, we often end up using cognitive processes that arose chiefly because of their role in making blame attributions.

To bring out the contrast between these two conclusions, we can return to the example of the accident and the investigator. Turning back to our example, once the investigator has finished figuring out why the accident occurred, he needs very little extra effort to figure out whether anyone is to blame. Solan believes that almost all of the processing needed to assess blame might already have been needed simply to figure out why the event occurred, with only a little bit of extra processing at the end being required

---

[3] *Id.* at 1009 (arguing that blame is triggered by a combination of the thought that an event occurred *because* of a person's action and that the person *should have known better*).

[4] *Id.* at 1004 (arguing that the impulse to blame is largely a 'by-product' of cognitive capacities we needed for other purposes); *id*. at 1012 (prescinding from any strong conclusions about the evolutionary basis of this outcome).

[5] *Id.* at 1004, 1012.

exclusively for the purpose of assessing blame.[6] By contrast, my conclusion is that the whole course of the investigator's work – even when he was only being asked to determine why the accident occurred – was shaped by a concern with issues of blame. The reason why so little additional processing is needed at the end is that, from the very beginning, his cognitive processes were shaped by a need to facilitate blame assessments.

In arguing for this conclusion, I focus on the two kinds of judgments that Solan discusses in his paper – judgments about mental states and judgments about causal relations.[7] My claim is that the way in which people make these judgments, even when they are not specifically being confronted with questions about blame, is deeply influenced by a concern with blame attributions.[8]

I. BLAME AND INTENTIONAL ACTION

Attributions of blame depend in a fundamental way on judgments about the agent's mental states.[9] Thus, our decision as to whether or not the agent is blameworthy will often depend on our judgments about that agent's goals, about the extent to which she foresaw certain outcomes, and about whether or not she performed the relevant behavior intentionally. But as Solan points out, we make these kinds of judgments all the time – even when we are not at all concerned with questions of blame – and it therefore appears that we use relatively little of the processing for which we detect mental states exclusively for the purpose of making blame assessments.[10]

A question then arises as to why we make these judgments in the way we do. One possible view would be that our capacity to detect and classify people's mental states arose, most fundamentally, from a need to predict and explain behavior. Then, given that we already had this capacity in place, we began using it in blame assessments as well.

---

[6] *Id.* at 1004.

[7] Solan, *supra* note 1, at 1014-20.

[8] *Id.* at 1018-20 (on mental states); *id.* at 1014-17 (on causal relations).

[9] Mark D. Alicke, *Culpable Control and the Psychology of Blame*, 126 PSYCHOL. BULL. 556, 566-68 (2000).

[10] Solan, *supra* note 1, at 1003.

But there is another possibility. Perhaps our capacity to detect and classify mental states has itself been shaped in certain ways by a need to assess blame. In other words, it might turn out that our capacity to detect mental states was not shaped *only* by a need for predictions and explanations, but also (at least in certain respects) by a need to determine whether or not particular agents are blameworthy.[11]

Take the distinction between behaviors that are performed "intentionally" and those that are performed "unintentionally." One hypothesis would be that this distinction was shaped primarily by a need for prediction and explanation. An alternative hypothesis would be that the distinction itself was shaped in part by a need for assessments of blame.

The best way to decide between these two hypotheses would be to look in detail at the criteria that people use when they are trying to figure out whether a given behavior was performed intentionally or unintentionally. Then we could see whether these criteria make better sense (a) as part of an attempt to predict and explain behavior or (b) as part of an attempt to assess blame. I have addressed this issue in a number of recent publications;[12] here we only have space for a highly compressed version of the argument.

When we want to investigate the criteria that people use in determining whether or not a behavior was performed intentionally, one of the most helpful methods is to look at people's intuitions regarding particular cases. For example, let us consider the following story:

> A lieutenant was talking with a sergeant. The lieutenant gave the order: 'Send your squad to the top of Thompson Hill.'
>
> The sergeant said: 'But if I send my squad to the top of Thompson Hill, we'll be moving the men directly into the enemy's line of fire. Some of them will surely be killed!'
>
> The lieutenant answered: 'Look, I know that they'll be in the line of fire, and I know that some of them will be killed. But I don't care at

---

[11] For a more radical view, see Kristin Andrews, *Folk Psychology is not a Predictive Device* (unpublished manuscript, on file with author) (arguing that our capacity to detect mental states was not shaped, even primarily, by need for prediction).

[12] *See, e.g.*, Joshua Knobe, *Intentional Action and Side Effects in Ordinary Language*, 63 ANALYSIS 190 (2003) [hereinafter Knobe, *Intentional Action and Side Effects*]; Joshua Knobe, *Intentional Action in Folk Psychology: An Experimental Investigation*, 16 PHIL. PSYCHOL. 309 (2003).

> all about what happens to our soldiers. All I care about is taking control of Thompson Hill.'
>
> The squad was sent to the top of Thompson Hill. As expected, the soldiers were moved into the enemy's line of fire, and some of them were killed.[13]

Confronted with this story, most people say that the lieutenant *intentionally* put the soldiers into the line of fire.

But suppose that we make a small change in the story, changing the effect of the lieutenant's behavior from something bad to something good. The story then becomes:

> A lieutenant was talking with a sergeant. The lieutenant gave the order: 'Send your squad to the top of Thompson Hill.'
>
> The sergeant said: 'If I send my squad to the top of Thompson Hill, we'll be taking the men out of the enemy's line of fire. They'll be rescued!'
>
> The lieutenant answered: 'Look, I know that we'll be taking them out of the line of fire, and I know that some of them would have been killed otherwise. But I don't care at all about what happens to our soldiers. All I care about is taking control of Thompson Hill.'
>
> The squad was sent to the top of Thompson Hill. As expected, the soldiers were taken out of the enemy's line of fire, and they thereby escaped getting killed.[14]

Confronted with this revised version of the story, most subjects actually say that the lieutenant did *not* intentionally take the soldiers out of the line of fire.[15] In fact, in a systematic experimental study, seventy-seven percent of subjects confronted with the first story said that the lieutenant intentionally put the soldiers into the line of fire, whereas only thirty percent of subjects confronted with the second story said that the lieutenant intentionally took the soldiers out of the line of fire.[16]

Results like these suggest that people actually use judgments about the goodness or badness of the outcome as part of the criteria by means of which they determine whether or not a given behavior was performed intentionally. But it seems unlikely that this aspect of the criteria serves primarily to facilitate some "scientific" purpose like the prediction and

---

[13] Knobe, *Intentional Action and Side Effects*, *supra* note 12, at 192.

[14] *Id.* at 192-93.

[15] *Id.* at 193.

[16] *Id*.

explanation of behavior. The most well-supported hypothesis (at least at this point in the evolving research on the topic) would be that the very criteria by means of which we distinguish between intentional and unintentional behaviors have been influenced in some way by a concern with issues of blame.

II. BLAME AND CAUSATION

Attributions of blame are influenced, not only by judgments about the agent's mental states, but also by judgments about causal relations.[17] In general, we are unlikely to blame the agent for an outcome unless we believe that the agent *caused* that outcome. But as Solan emphasizes, people quite often try to figure out whether or not a particular agent caused a particular outcome even when they are not wondering whether or not the agent is to blame.[18] After all, a proper understanding of causal relations is often helpful in predicting and explaining events.

This is quite a striking fact. It seems odd that the very same relation – the relation of causation – should be used both for assessing blame and for generating predictions and explanations. Why don't we use two different relations here – one relation for assessing blame and another, slightly different relation for prediction and explanation? Solan is careful not to engage in dogmatic evolutionary speculation. However, he does suggest an interesting possibility. Perhaps we already needed a capacity for detecting causal relations (because this capacity was useful in generating predictions and explanations), and we then came to use this capacity for assessing blame as well.[19] But here again, there is another possibility. Perhaps our capacity for detecting causal relations was itself shaped in a fundamental way by our concern with questions of blame.

Note that we are not here entertaining the absurd hypothesis that people's whole capacity for detecting causal relations arose out of a need to make assessments of blame. The idea is simply that certain aspects of this capacity – a capacity that presumably arose chiefly out of a need for prediction and explanation – may also have been shaped by a

---

[17] Solan, *supra* note 1, at 1004.

[18] *Id.*

[19] *Id.* at 1004, 1012.

concern with attributions of blame. To test this idea, we can look closely at the criteria by which people decide whether or not a given agent was the cause of a given outcome. The question is whether all aspects of these criteria can be understood as part of an attempt to arrive at accurate predictions and explanations or whether some aspects only make sense as part of an attempt to assess blame.

In this connection, let us consider the following story:

> Lauren works in the maintenance department of a large factory. It is her responsibility to put oil in the K4 machine on the first day of each month. If she doesn't put in the oil, the machine will break down.
>
> On June first, Lauren forgot to put in the oil. The machine broke down a few days later.

Here it seems at least somewhat natural to say that Lauren caused the machine to break down. After all, if she had simply fulfilled her responsibility and put in the oil, the breakdown would never have occurred.

But now suppose that we add a new character to our story:

> Jane also works in the factory, but she does not work in the maintenance department. She works in human resources, keeping track of all the details for the employee health insurance plan.
>
> Jane also knew how to put oil in the K4 machine. But no one would have expected her to do so; it clearly wasn't part of her job.

Although Jane is quite similar to Lauren in certain respects, it seems quite wrong to say that Jane caused the accident. Indeed, I conducted a simple experiment to show that people are more inclined to think that Lauren caused the accident than that Jane caused it.[20]

But why do we distinguish between Lauren and Jane in this way? Neither of them put oil into the machine, and if

---

[20] The subjects of this study were thirty-five people spending time in a Manhattan public park. All of the subjects received the same questionnaire. First, they read the vignette about Lauren, followed by the question: "Did Lauren *cause* the machine to break down?" Then they were asked to read the vignette about Jane, followed by the question: "Did Jane *cause* the machine to break down?" Each question was answered on a scale from zero ("no, she didn't") to six ("yes, she did"). The mean rating for the Lauren vignette (M=.37) was significantly lower than that for the Jane vignette (M=3.34), $t(35)=7.2$, $p<.001$. In other words, the degree to which people thought that Lauren was the cause was so much lower than the degree to which people thought that Jane was the cause that the difference is extremely unlikely to be due to chance alone.

either of them had put the oil in, the machine would not have broken down. Why then do we say that Lauren caused the breakdown but Jane did not? In cases like this one, it seems hard to deny that our judgments about causal relations are being influenced in some way by our beliefs about the rightness and wrongness of particular behaviors.[21] Presumably, we are influenced by the thought that Lauren was doing something *wrong*, that she really *shouldn't* have neglected to put oil in the machine.

What we see here, apparently, is a sense in which our capacity to detect causal relations is sensitive to moral considerations. But it seems unlikely that this sensitivity is somehow furthering our aim of generating accurate predictions and explanations. Thus, although these phenomena are not yet well-understood, it seems that the balance of evidence now points to the view that our capacity to detect causal relations has been shaped in certain respects by a concern with issues of blame.

III. CONCLUSION

Solan has directed our attention to an extremely important phenomenon: The surprising overlap between the cognitive capacities that we use when assessing blame and the capacities that we use for other, unrelated purposes.[22] It appears that the vast majority of the capacities that we use when assessing blame are also used when we are simply trying to figure out why some given event has occurred.[23]

Drawing upon this phenomenon, Solan is able to provide some enticing evidence for the conclusion that our concern with blame has had a relatively small impact on our underlying cognitive capacities.[24] The essence of his argument lies in the claim that, since we already needed so many of the relevant capacities for other purposes, only a relatively small amount of additional structure would be necessary to make possible the ability that we now have to assess blame.[25]

---

[21] For similar views, see generally Judith Jarvis Thomson, *Causation: Omissions*, 66 PHIL. & PHENOMENOLOGICAL RES. 81 (2003); Sarah McGrath, *Causation by Omission: A Dilemma*, 123 PHIL. STUD. 125 (2005).

[22] Solan, *supra* note 1, at 1004.

[23] *Id.*

[24] *Id.* at 1004, 1012.

[25] *Id.* at 1004.

Although future research may vindicate Solan's argument, it seems to me that the presently-available research actually points more strongly to the opposite conclusion. It is true that most of the capacities that we use when assessing blame are also used when we are simply trying to figure out why an event occurred. But we should not therefore assume that those capacities were already needed for some other purpose and then came to be used in blame assessment as well. Another possible conclusion – and one for which I have presented some tentative support – is that the capacities we normally use to explain and interpret events have been shaped in a fundamental way by our concern with blame.