Cognitive Processes Shaped by the Impulse to Blame

Joshua Knobe

Princeton University

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 processes that we sometimes use even when we aren't even considering the issue of blame. 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, 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

[1] Lawrence Solan, *Cognitive Foundations of the Impulse of Blame*. BROOKLYN LAW REVIEW (forthcoming).

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.

Solan provides support for this initial intuition through a sophisticated analysis of the cognitive processes that underlie attributions of blame. He shows that attributions of blame rely in a crucial way on judgements about *mental states* and about *causal relations*. Then he shows that we would have made these very judgements anyway, even if we hadn't 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.

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 almost 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. I would draw the opposite conclusion: that blame has had such a pervasive influence on our cognitive capacities that, even when we aren't 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. The phenomenon is that, 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's conclusion is that almost all of the processing needed to assess blame was already needed simply to figure out why the event occurred, with only a little bit of extra processing at the end being required exclusively for the purpose of assessing blame. 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 judgements that Solan discusses in his paper — judgements about mental states and judgements about causal relations. My claim will be that the way in which people make these judgements, even when they are not specifically being confronted with questions about blame, is deeply influenced by a concern with blame attributions.

I

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

A question then arises as to why we make these judgements in the way we do. Take the distinction we normally draw between 'intentional' and 'unintentional' behaviors. Why exactly do we make this distinction? One possibility would be that we make this distinction because, by making it, we are able to do a better job of predicting and explaining behavior. Then, given that we already had the distinction in place, we have come to use it in blame assessments as well

But there is also another possibility. Perhaps the distinction itself has been shaped in part by our concern with issues of blame. We would then be left with a quite different picture of the relationship between our use of this distinction in assessing blame and our use of the distinction for various other purposes. The picture would not be that we already needed the distinction for other purposes and then started using it to assess blame. Rather, it would be that part of the reason why we sometimes use this distinction for other purposes is that it was extremely important in the context of blame assessment and therefore came to be a central aspect of the framework by means of which we classify and interpret behavior.

The best way to test this latter hypothesis 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;[2] 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:

[2] See Joshua Knobe, *Intentional Action and Side Effects in Ordinary Language*, ANALYSIS (forthcoming), as well as Joshua Knobe, *Intentional Action in Folk Psychology: An Experimental Investigation*, PHILOSOPHICAL PSYCHOLOGY (forthcoming).

> 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 the 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 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.

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.

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. In fact, in a systematic experimental study, 77% of subjects confronted with the first story said that the lieutenant intentionally put the soldiers into the line of fire, whereas only 30% of subjects confronted with the second story said that the lieutenant intentionally took the soldiers out of the line of fire.

What results like this one suggest is that people actually use judgements 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 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

Attributions of blame are influenced, not only by judgements about the agent's mental states, but also by judgements about causal relations. 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 aren't wondering whether or not the agent is to blame. 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's answer is that we already needed a capacity for detecting causal relations (because this capacity was useful in generating predictions and explanations) and that we then came to use this capacity for assessing blame as well. 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 predictions and explanations — may also have been shaped by a 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 will be 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 1st, Lauren forgot to put in the oil. The machine broke down a few days later.

Here it seems 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 suppose now 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 work 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 here quite similar to Lauren in certain respects, it seems quite wrong to say that Jane caused the accident.

But why do we distinguish between Lauren and Jane in this way? Neither of them put oil into the machine, and if 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 and Jane did not? In cases like this one, it seems hard to deny that our judgements about causal relations are being influenced in some way by our beliefs about the rightness and wrongness of particular behaviors. 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

Solan has directed our attention to an extremely important phenomenon. This phenomenon is the surprising overlap between the cognitive capacities that we use when assessing blame and the capacities that we use for other, unrelated purposes. 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.

Drawing upon this phenomenon, Solan argues for the conclusion that our concern with blame has had a relatively small impact on our underlying cognitive capacities. 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.

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.