THE JOURNAL OF PHILOSOPHY
VOLUME CVI, NO. 11, NOVEMBER 2009

CAUSE AND NORM*

Much of the philosophical literature on causation has focused on the concept of "actual" causation, sometimes called "token" causation. In particular, it is this notion of actual causation that many philosophical theories of causation have attempted to capture.[1] In this paper, we address the question: What purpose does this concept serve? As we shall see in the next section, one does not need this concept for purposes of prediction or rational deliberation. What then could its purpose be? We will argue that one can gain an important clue here by looking at the ways in which causal judgments are shaped by people's understanding of *norms*.

* We would like to thank Nancy Cartwright, Clark Glymour, Alison Gopnik, Dennis Hilton, Christoph Hoerl, David Lagnado, Tania Lombrozo, David Mandel, Laurie Paul, Jonathan Schaffer, Jim Woodward, Gideon Yaffe, and audience members at the McDonnell workshop on Causal and Moral Cognition (California Institute of Technology), the University of Southern California, the workshop on the Origins and Functions of Causal Thinking IV (Venice, Italy), Rutgers University, the Society for Philosophy and Psychology (Toronto), the Workshop on Causal and Counterfactual Understanding (University of Warwick), and the Workshop on Counterfactuals (Erasmus University, Rotterdam).

[1] To name just a few: Donald Davidson, "Causal Relations," in *Essays on Actions and Events* (New York: Oxford, 1980), pp. 149–62; Phil Dowe, *Physical Causation* (New York: Cambridge, 2000); Ellery Eells, *Probabilistic Causality* (New York: Cambridge, 1991), chapter 6; Joseph Halpern and Judea Pearl, "Causes and Explanations: A Structural-model Approach—Part I: Causes," *British Journal for the Philosophy of Science*, LVI (2005): 843–87; Christopher Hitchcock, "The Intransitivity of Causation Revealed in Equations and Graphs," this JOURNAL, XCVIII, 6 (June 2001): 273–99; "Prevention, Preemption, and the Principle of Sufficient Reason," *Philosophical Review*, CXVI (2007): 495–532; David K. Lewis, "Causation," this JOURNAL, LXX, 17 (October 11, 1973): 556–67, reprinted with postscripts in *Philosophical Papers, Volume II* (New York: Oxford, 1986), pp. 159–213; "Causation as Influence," this JOURNAL, XCVII, 4 (April 2000): 182–97; Pearl, *Causality: Models, Reasoning, and Inference* (New York: Cambridge, 2000), chapter 10; Michael Tooley, *Causation: A Realist Approach* (New York: Oxford, 1987); James Woodward, *Making Things Happen* (New York: Oxford, 2003), chapter 2; Stephen Yablo, "De Facto Dependence," this JOURNAL, XCIX, 3 (March 2002): 130–48. See also the many essays in John Collins, Ned Hall, and L.A. Paul, eds. *Causation and Counterfactuals* (Cambridge: MIT, 2004), and Dowe and Paul Noordhof, eds., *Cause and Chance* (New York: Routledge, 2004).

0022-362X/09/0611/587–612

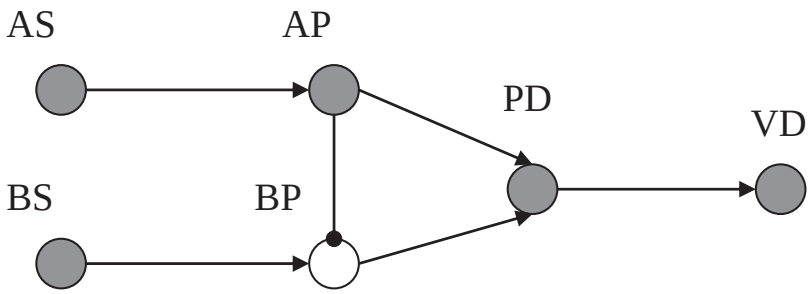

Figure 1

I. THE PROBLEM

We may illustrate the concept of actual causation with a simple example:

> Assassin and Backup set off on a mission to poison Victim. Assassin puts poison in Victim's drink. Backup stands at the ready; if Assassin hadn't poisoned the drink, Backup would have. Both poisons are lethal. Victim drinks the poison and dies.

This is an example of causal *preemption*: Assassin's action causes Victim's death, and also preempts Backup's action, which would have caused the death if Assassin had not acted. Cases of causal preemption have received a lot of attention, since they provide problems for both regularity and counterfactual theories of causation. Backup's presence, plus his willingness to use the poison, together with the composition of the poison, Victim's thirst, and other factors, are nomically sufficient for Victim's death without the need to mention Assassin; nonetheless, Backup is not a cause of Victim's death. And Victim's death would have occurred even if Assassin had not acted; nonetheless, Assassin's action is a cause of Victim's death.

Our little story has a *causal structure*, which can be represented abstractly using a neuron diagram (figure 1) or a system of structural equations as follows:

$$\begin{aligned} AS &= 1 \\ BS &= 1 \\ AP &= AS \\ BP &= BS \,\&\, \sim AP \\ PD &= AP \vee BP \\ VD &= PD \end{aligned}$$

(Where AS represents 'Assassin sets off', BS 'Backup sets off', AP 'Assassin poisons the drink', BP 'Backup poisons the drink', PD 'The drink is poisoned', and VD 'Victim dies'.) These representations tell us how each event in the story depends upon the other events in the story. For

example, they tell us that Backup would have poisoned the drink just in case he set off and Assassin had not poisoned the drink. Note that there is nothing inherently general or universal about the causal structure. For example, the last equation only tells us that in this particular case, Victim would have died if his drink had been poisoned and would not have died if it had not been poisoned. It does not say that poisonings always, or even typically, cause deaths.[2]

Over and above this causal structure, however, we have the judgment that Assassin's actions *actually caused* Victim to die, while Backup's actions did not. This is a judgment about actual causation. We wish to ask what the purpose of such a judgment is. In particular, why don't we make do with just the causal structure? After all, the causal structure suffices to allow us to make predictions about what will happen, given any combination of the relevant causal antecedents. The causal structure allows us to infer what would happen if we were to interfere with the system to disrupt the causal relationships in various ways. And the causal structure allows us to answer any counterfactual questions about how things might have gone differently. What more could we want from knowledge of causation?

We may approach this problem from a slightly different direction. There have been a number of attempts to extend the counterfactual theory of causation to accommodate cases of causal preemption. One promising line, realized in somewhat different ways by David Lewis,[3] Joseph Halpern and Judea Pearl,[4] Ned Hall,[5] and Christopher Hitchcock,[6] is to identify causation not with counterfactual dependence in the actual situation, but rather with counterfactual dependence in a certain kind of "normalized" version of the actual situation. This normalized situation is reached by replacing abnormal features of the actual situation with more normal alternatives. Thus in our little causal story, while Victim's death does not counterfactually depend upon Assassin's action in the actual situation described, it would depend upon Assassin's action in the normalized version of the situation where Backup is not present. This is a rough sketch only; the details need not detain us. The question that arises is: What purpose can be served by a concept that has these contours? What purpose can be

[2] See Hitchcock, "Three Concepts of Causation," *Philosophy Compass*, II (2007): 508–16 (http://www.blackwell-synergy.com/toc/phco/2/3), and "Prevention, Preemption, and the Principle of Sufficient Reason," for further discussion of this point.

[3] Lewis, Postscript E to "Causation," in his *Philosophical Papers: Volume II*, pp. 193–213.

[4] Halpern and Pearl, "Causes and Explanations."

[5] Hall, "Structural Equations and Causation," *Philosophical Studies*, CXXXII (2007): 109–36.

[6] Hitchcock, "Prevention, Preemption, and the Principle of Sufficient Reason."

served by identifying factors that would have made a difference, not in the actual situation, but in a modified version of the actual situation?

II. SOLUTION SKETCH

Before we begin examining the purpose of the concept of actual causation, it might be helpful to consider one popular view about the purpose of the concept of causal structure. Several authors have championed the view that the purpose of understanding causal structure is to predict the outcomes of *interventions*.[7] In particular, it has been suggested that the key difference between knowing causal structure and merely knowing facts about correlations is that the former type of knowledge can enable people to predict the outcomes of intervention in a way that the latter cannot.

To see the distinctive role of the concept of causation here, consider what might happen if you discovered a correlation whereby people who are beaten as children tend to be more violent as adults. Just by knowing that correlation, you could predict that a child who was beaten would be more likely to later become violent. Why then would you want to know whether there was truly a *causal* relationship between being beaten and becoming violent? The idea we will be drawing on here is that this information about causation proves useful when one's goal is not only to predict the results of subsequent observations but also to actively intervene in the world. Hence, in this example, the important thing about knowing whether there was a causal relationship here would be that this sort of knowledge would allow you to determine whether you could prevent people from becoming violent in adulthood by intervening to prevent them from being beaten as children.

Our aim here is to construct an account that extends this basic insight, showing how people's concept of actual causation enables them to design effective interventions. In particular, we argue that information that has nothing to do with causal structure can sometimes prove helpful in determining which intervention would be most suitable and that people are actually making use of this information in deciding which factors to denominate as "causes." In general, while causal structure identifies all of the factors that *could* be manipulated (either singly or in combination) to effect a change in the outcome, the actual causes are the factors that *should* be manipulated.

For a simple example, consider a student who has just gotten an F on a test and is wondering how this outcome might have been avoided. At

[7] For example, Nancy Cartwright, "Causal Laws and Effective Strategies," *Noûs*, XIII (1979): 419–37; Woodward, *Making Things Happen*.

least in theory, she might consider all sorts of different possibilities: "I would not have gotten an F if the teacher had been eaten by a lion." "I would not have gotten an F if the Earth's gravitational pull had suddenly decreased." "I would not have gotten an F if I had had less to drink the night before the test." Yet although all of these different counterfactuals might be correct, she is only likely to entertain one of these, and only one of them actually points toward a suitable target of intervention. She should not be wasting her time thinking about the possibility of avoiding a bad grade by introducing lions or decreasing the Earth's gravitational pull. The thing to focus on here is the possibility of getting a better grade by drinking less.

It is in addressing problems like this one, we think, that the concept of actual causation really earns its keep. Information about causal structure entails many different counterfactuals about how a given outcome might have been avoided. But people's judgments of actual causation do something more. They select, from amongst all of these different counterfactuals, a chosen few that might actually be worth considering. In short, the concept of actual causation enables us to pick out appropriate targets for intervention.[8]

At this point, it might be helpful to clarify the status of the claim we are making. We certainly are not saying that people's ordinary criteria for actual causation include something like: "Pick out the factor that would be the most appropriate target for intervention." Nor are we claiming to offer anything like an analysis of actual causation. Rather, the significance of questions about targets of intervention only arises when we ask a further justificatory question. Even if we were able to develop a perfect set of criteria for actual causation, there would still be a further question, namely, *Why deploy a concept with just these criteria?* It is here that the notion of identifying suitable targets of intervention plays a role. We will argue that there truly is something of value about the criteria people ordinarily use, and the claim will be that this value comes from the fact that people's ordinary criteria tend to pick out factors that would genuinely be suitable targets of intervention.

The basic strategy we find here is much like the one at work in interventionist theories of causal structure. As we saw above, the motivating

[8] Our account has affinities with the proposal—championed by Guido Calabresi in *The Costs of Accidents* (New Haven: Yale, 1970)—that the agent who caused some harmful outcome is the one who could have avoided the outcome with the least cost. Under this concept of causation, it is argued, the practice of holding those who cause bad outcomes responsible for those outcomes will maximize social utility; those who can avoid bad outcomes most cheaply will be provided with an incentive for paying those costs, and in this way, the cost to the society as a whole is minimized. As we shall see in section IX below, however, our evaluative criteria have nothing to do with economic cost.

idea behind these accounts is that causal relationships, in contrast with spurious correlations, for example, can support our interventions. But this idea, by itself, is clearly inadequate as an analysis of the distinction between causal relationships and mere correlations. For example, there can be causal relationships in astrophysics, where human intervention is impossible in practice or even in principle. The actual criteria for causation do not themselves make any reference to agents trying to achieve goals; instead, they are framed in terms of certain technical concepts which need not detain us here.[9] But then one can still ask why it might make sense to employ a concept with these particular criteria. Here the suggestion is that these criteria make sense because they generally enable us to pick out factors that we can manipulate in order to achieve particular goals. The appeal to agency is not part of the analysis of causation, but rather helps to explain why we put so much stock in causal knowledge.

Our approach here has the same basic structure. First we point to some surprising features of the actual criteria governing ascriptions of causation. Then we suggest that these criteria actually make sense if one assumes that causal ascriptions serve a particular kind of purpose in our lives. It is only at this second level that we invoke the idea of identifying suitable targets for intervention.

III. CAUSAL SELECTION

It has long been recognized that ordinary causal judgments make use of information that goes beyond anything that might be included in causal structure. People seem to rely on extra-structural information to *select* certain candidate causes over others—claiming, for example, that the spark was a "cause" of the fire, while the oxygen was merely a background "condition."[10] Yet although everyone acknowledges that this process of selection takes place, questions about the precise nature of the selection have not figured prominently in the existing literature on the concept of causation. Indeed, most authors who have chosen to discuss the problem of selection at all have ended up concluding that it serves only to distract us from those aspects of causal judgment that are truly philosophically important. Thus Mill writes:

> It is seldom, if ever, between a consequent and a single antecedent, that this invariable sequence subsists. It is usually between a consequent and the sum of several antecedents; the concurrence of all of them being

[9] See for example, Woodward, chapters 2 and 3.

[10] See, for example, H.L.A. Hart and T. Honoré, *Causation in the Law*, Second Edition (New York: Oxford, 1985), especially pp. 32–41.

> requisite to produce, that is, to be certain of being followed by, the consequent. In such cases it is very common to single out one only of the antecedents under the denomination of Cause, calling the others mere Conditions .... The real cause, is the whole of these antecedents; and we have, philosophically speaking, no right to give the name of cause to one of them, exclusively of the others.[11]

Lewis says:

> We sometimes single out one among all the causes of some event and call it 'the' cause, as if there were no others. Or we single out a few as the 'causes', calling the rest mere 'causal factors' or 'causal conditions' .... I have nothing to say about these principles of invidious discrimination.[12]

And Hall adds:

> When delineating the causes of some given event, we typically make what are, from the present perspective, invidious distinctions, ignoring perfectly good causes because they are not sufficiently salient. We say that the *lightning bolt* caused the forest fire, failing to mention the contribution of the oxygen in the air, or the presence of a sufficient quantity of flammable material. But in the egalitarian sense of 'cause', a complete inventory of the fire's causes must include the presence of oxygen and of dry wood.[13]

In short, all of these authors agree that our ordinary causal attributions are highly selective, but they nonetheless contend that this selectivity is somehow independent of the truly deep and important features of the concept of causation. These deep and important features, they argue, are entirely "egalitarian," that is, free from the sort of selectivity one finds in ordinary causal attributions.

We concur that this egalitarianism is entirely appropriate at the level of what we have called causal structure.[14] We claim, however, that it is out of place at the level of actual causation. If the concept of actual causation were entirely egalitarian, we find it hard to see how it could be helpful for people even to have the concept at all. All of the truly important information would already be contained in the causal structure, and it seems that people would be better off ignoring questions of actual causation altogether and simply talking about patterns of counterfactual dependence. The value of the concept of actual causation, we wish to suggest, comes precisely from the fact that it is *inegal-*

[11] Mill, *A System of Logic, Volume I*, Fourth Edition (London: J.W. Parker and Son, 1856), pp. 360–61.

[12] Lewis, "Causation," cited from *Philosophical Papers*, p. 162.

[13] Hall, "Two Concepts of Causation," in Collins, Hall, and Paul, eds., p. 228.

[14] See Hitchcock, "Three Concepts of Causation," for further discussion.

itarian. Our concept of actual causation enables us to pick out those factors that are particularly suitable as targets of intervention.

IV. NORMS AND CAUSAL JUDGMENT

To illustrate the basic issues here, we can introduce a simple case, which we will call the *pen vignette*:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take 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 repeatedly e-mails them reminders that only administrators 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, if these events were actually unfolding, there would be a number of different ways of intervening to prevent the problem. One approach would be to make the professor do what he was supposed to do and stop violating the rules. Another would be to allow the professor to violate the rules but to cancel out some of the effects of his action by making the administrative assistant refrain from taking any pens. Either of these two approaches would work in the situation as described. And yet, it seems that the two approaches are not entirely equal. One has a clear intuition—an intuition we will explore at length below—that it would actually be more suitable in some important respect to intervene by making the professor obey the rules.

The key thing to note now is that people's ordinary causal judgments follow this same pattern. In one recent experiment, subjects were given the pen vignette and then asked whether they agreed or disagreed with one of the following sentences.

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

Overall, subjects agreed with the statement that Professor Smith caused the problem, but they disagreed with the statement that the administrative assistant caused the problem.[15] In other words, the factor that

[15] J. Knobe and B. Fraser, "Causal Judgment and Moral Judgment: Two Experiments," in Walter Sinnott-Armstrong, ed., *Moral Psychology, Volume 2: The Cognitive Science of Morality* (Cambridge: MIT, 2008), pp. 441–48. Subjects were presented with one of the two

subjects picked out as a "cause" was precisely the factor that seemed most suitable as a target of intervention.

This simple example illustrates the basic elements of the account of causal judgment we will be developing here. On this account, people's causal judgments do not simply pick out any old factor that happens to stand in the right position within a causal structure. Instead, ordinary causal judgments specifically pick out those factors that would prove especially suitable as targets of intervention. It is this selective aspect of causal judgments that allows them to have a value that goes beyond anything one might find in causal structure alone.

Our argument for this account comes in two parts. First we engage in a purely descriptive project, trying to understand the pattern of people's ordinary causal judgments. Then we turn to more explicitly evaluative questions, explaining how the pattern evinced in people's ordinary judgments might actually prove helpful in selecting targets of intervention.

V. COUNTERFACTUAL REASONING AND NORMS

To arrive at a proper understanding of people's ordinary causal judgments, we will need to turn to a topic that might at first seem unrelated to the issue of causal selection but which, we believe, will eventually help us to resolve the problems under discussion here. In particular, we need to turn to questions about the human capacity for *counterfactual reasoning*—the capacity people use for thinking about how things might have been different from the way they actually are.

Our concern here will not be with the aspects of counterfactual reasoning that usually come up in discussions of topics like this one. We will not be asking about the semantics of counterfactuals or about how people determine whether certain counterfactuals are true or false. Instead, we will be concerned with questions about how people decide which counterfactuals are worth thinking about in the first place. Or, to put it in the terms we will be using here, our concern is with questions about how people decide which counterfactuals are *relevant*. These questions have not played much of a role in the philosophical literature thus far, but they have been investigated in great depth

---

statements, and were asked to rate the extent to which they agreed or disagreed with it on a 7-point scale, from 1 (strongly disagree) to 7 (strongly agree). Those who were given the statement 'Professor Smith caused the problem' gave an average rating of 5.2, while those who were given the statement 'The administrative assistant caused the problem' gave an average rating of 2.8. This difference was statistically significant ($p < .001$).

within cognitive and social psychology,[16] and quite a bit is now known about the ways in which people determine whether or not a given counterfactual is relevant.

For an example of the phenomenon under discussion here, consider what might happen if we submitted a manuscript to a journal and it ended up being sent to a reviewer who held certain bizarre and idiosyncratic views. To make the case more concrete, let us suppose that the reviewer has a strong distaste for the word 'and' and that he therefore rejects our paper on the grounds that one should never use that word more than three times per page.[17] It seems that we would then treat as relevant counterfactuals of the form:

(1) If the manuscript had been sent to a reviewer with more ordinary views ....

Thoughts about this sort of counterfactual would leap immediately to our minds; we might even be unable to keep ourselves from thinking about them. By contrast, we would probably treat as completely irrelevant counterfactuals of the form:

(2) If the reviewer had composed a catchy pop song about our manuscript ....

Of course, people are not literally *incapable* of thinking about irrelevant counterfactuals like this one. (If pressed, they could imagine the counterfactual situation and do their best to figure out what outcomes would result.) Still, there is an almost overwhelming tendency for people not to think in any way about such things. Our capacity for counterfactual reasoning seems to show a strong resistance to any consideration of irrelevant counterfactuals.

Ultimately, our aim is to show that certain facts about the way people assess the relevance of counterfactuals can explain the puzzling patterns we observed earlier in looking at the nature of causal selection. That is, we want to show that the criteria people use when assessing relevance can be used to explain how people manage to pick

[16] See, for example: Ruth Byrne, "Mental Models and Counterfactual Thoughts about What Might Have Been," *Trends in Cognitive Science*, VI (2002): 426–31; Daniel Kahneman and Dale Miller, "Norm Theory: Comparing Reality to Its Alternatives," *Psychological Review*, LXXX (1986): 136–53; Kahneman and Amos Tversky, "The Simulation Heuristic," in Kahneman, Paul Slovic, and Tversky, *Judgment under Uncertainty: Heuristics and Biases* (New York: Cambridge, 1982), pp. 201–10; David R. Mandel, Denis J. Hilton, and Patrizia Catellani, eds., *The Psychology of Counterfactual Thinking* (New York: Routledge, 2005); N.J. Roese, "Counterfactual Thinking," *Psychological Bulletin*, CXXI (1997): 133–48.

[17] One of us in fact had a high school philosophy teacher with just this idiosyncrasy; that is the reason for the excessive use of colons, semicolons, em dashes, and similar and-avoidance strategies in the present paper.

out the most suitable targets of intervention. For the moment, however, we want to put aside these broader issues about the relationship between causation and moral judgment and simply focus on developing a theory about how people assess the relevance of counterfactuals.

At the heart of our theory is the idea that people's judgments about the relevance of counterfactuals depend in an essential way on *norms*. The basic suggestion is that people classify events on a scale from "normal" to "abnormal." Then, when something abnormal occurs, they regard as relevant counterfactuals those that involve something more normal having occurred instead. Roughly speaking, the reason why people think that counterfactual (1) is relevant is that they think it is *normal* for philosophy papers to be given to reviewers who do not evaluate papers based on the frequency of an innocuous conjunction. Conversely, the reason people regard counterfactual (2) as irrelevant is that they think it is *abnormal* for reviewers to write catchy pop songs about the manuscripts they receive. And similarly for numerous other sorts of cases. The basic picture here is one in which, whenever something abnormal occurs, people show a tendency to think about how things would have gone if something normal had taken place instead. By contrast, people are much less inclined to entertain counterfactual hypotheses in which normal events are replaced by abnormal ones.

At this point, however, it may be thought that our account suffers from a fundamental ambiguity. For it might be suggested that terms like 'norm' and 'normal' can be used in a number of distinct senses. Indeed, one might say that there are at least two completely separate meanings at work here.

First, there are what one might call "statistical norms." These norms simply capture information about the relative frequencies of certain events. Using the word 'norm' in this sense, one might say, for example: "The temperature yesterday was considerably colder than the norm for this time of year."

Second, there are "prescriptive norms"—norms that actually tell us what ought to happen under certain circumstances. Using 'norm' in this latter sense, one might say: "There is a strong norm in our community against interrupting during someone's talk." Moreover, there seem to be different kinds of prescriptive norms. There are purely moral norms, where violating the norm would be intrinsically wrong. Then there are legal norms, or more generally, norms arising from policies adopted by social institutions. Professor Smith violated such a norm when he took one of the last pens. Often there is nothing *intrinsically* wrong with an action that violates such a norm, although the presence of the policy (if just and conducive toward some social good) might create a moral obligation to adhere to it. There are also

norms of proper functioning that apply to artifacts and biological organisms (and their components).

In light of these distinctions, it might be thought that we need to say which specific type of norm ends up playing a role in the way people assess the relevance of counterfactuals. Yet although there clearly are differences between these kinds of norms, it also seems that it is not merely a sort of pun that they have come to be denoted by the same word. There really does seem to be some important way in which these different kinds of norms are connected in our ordinary ways of understanding the world. Hence, when we say of a person that she is "abnormal," we do not typically have in mind either a purely statistical judgment or a purely prescriptive one. Instead, we seem to be making a single overall judgment that takes both statistical and prescriptive considerations into account.

Perhaps the connection we see here between statistical and prescriptive norms stems from the fact that people can use each of these norms as a heuristic for the other. On the one hand, when we are trying to figure out what people are going to do, we sometimes rely on the assumption that they will probably end up choosing the option that is best. On the other, when we are wondering whether a given option is a bad one, we sometimes proceed by assuming that an option is unlikely to be very bad if it is frequently chosen. Given the close connection between these two kinds of judgments, people may have developed a single type of representation that is not specifically designed either for statistical purposes or for prescriptive purposes but which manages to do a fairly decent job in both domains. To take just one example, it seems that people sometimes develop "scripts" for stereotypical situations and that, although these scripts do not perfectly capture either the statistical distribution of behaviors in those situations or the prescriptively optimal behaviors to perform, they do help people to get a rough sort of handle on both of these issues.

In keeping with this general idea, our theory does not assign distinct roles to statistical and prescriptive norms. The claim is rather that people classify each counterfactual as having a single overall degree of "normality" (taking both statistical and prescriptive considerations into account) and that this overall degree of normality ends up affecting people's sense of whether the counterfactual is relevant or irrelevant.

With this framework in place, we can return to the example discussed above. The suggestion would be that people develop an overall conception of what sorts of criteria for evaluating a paper would be "normal" for a reviewer to apply. This conception integrates both statistical and prescriptive considerations—picking out a particular class of criteria in a way that generally steers clear of those that are infre-

quent (the statistical considerations) but also, so far as possible, steers clear of criteria that are not appropriate for evaluating philosophy papers (the prescriptive considerations). People's counterfactual reasoning then gravitates toward thoughts of reviewers who apply criteria that are normal in this sense. Thus, when the actual reviewer's views are abnormal, people tend to think about what would have happened if they had been more normal.

Looking through the psychological literature, one finds a number of different hypotheses about precisely how this reasoning process takes place,[18] but the precise details of these hypotheses will not be important here. The key point for present purposes is just that, whatever else might be going on, people do generally show a tendency to direct their counterfactual thoughts toward ways in which the world could have been more normal—and there is considerable empirical support for this general claim. To take a case involving prescriptive norms: Read[19] presented subjects with a vignette in which two people are playing cards. One person wins the hand, and subjects are invited to complete the sentence "The outcome would have been different if …." Subjects were much more strongly inclined to strengthen the losing hand than to weaken the winning hand. Or, for a case involving statistical norms: Kahneman and Tversky[20] presented subjects with a vignette in which Mr. Jones was killed in a traffic accident on the way home from work when he was hit by a drug-crazed truck-driving teenager who ran a red light. In one version of the story, Mr. Jones left home from work early in order to run some errands; in another version, he deviated from his normal route home in order to enjoy the scenic drive along the lake. Subjects were informed of both possibilities; for example, subjects who were given the first version of the story were also told that Mr. Jones occasionally took a scenic route home, but that he took his normal route home that day. Subjects were then told that Jones's family and friends often found themselves thinking "if only…"; the subjects were then told to complete the "if only…" statements. Those subjects who were given the first version of the story were strongly inclined to say "if only Jones had left at his regular time," rather than "if only Jones had taken the scenic route"; whereas subjects who were given the second version of the story were inclined to say "if only Jones had taken his regular route home" rather than "if only Jones had left early to run errands." And there are numerous further studies showing the same basic effect.

[18] For example, Kahneman and Miller, *op. cit.*; Roese, *op. cit.*
[19] D. Read, "Deteminants of Relative Mutability," unpublished manuscript.
[20] Kahneman and Tversky, *op. cit.*

VI. EVALUATING THE CRITERIA FOR COUNTERFACTUAL RELEVANCE

Moreover, when one examines these phenomena from a more evaluative perspective, one sees immediately that they really do make a certain amount of sense. It is not just that our minds are constructed in such a way that we tend to think about what would have happened if the reviewer had applied more ordinary criteria—it truly is *helpful* to think about counterfactual scenarios like that one. Similarly, it is not just that people happen not to think about what would have occurred if the reviewer had decided to write a pop song—it truly would be *pointless* to waste one's time on thoughts like that one. The best way to see why this is so is to look separately at the significance of statistical norms and prescriptive norms. In both cases, however, the key points will be remarkably straightforward.

The reason why it is helpful to think about possibilities that are statistically frequent is that these possibilities are the ones that we are most likely to encounter in the future. In one particular instance, we may happen to encounter a highly unusual sort of situation, but it would be a mistake in such cases simply to focus on the details of what actually ended up occurring. If we want to plan adequately for the future, we need also to think about what would have happened if a more probable type of situation had arisen.

The reason why it is helpful to think about possibilities in which prescriptive norms are not violated is that those are typically the possibilities we are trying to bring about. If we are trying, for example, to find some way to avoid a particular outcome, we do not want to search for just *any* way of avoiding that outcome. Rather, we want to find a way of avoiding the outcome without also having something else go wrong. (Moreover, when we do not have accurate statistical information available, we use conformity to a prescriptive norm as a heuristically useful surrogate.)

All in all, then, it seems fairly easy to see how it might be helpful for people to engage in counterfactual reasoning in the way they do. What we want to show now is that people's tendency to consider certain counterfactuals rather than others can actually explain the puzzling patterns we find in their causal intuitions.

VII. CAUSAL JUDGMENTS AND RELEVANT COUNTERFACTUALS

Since the publication of Lewis's "Causation," there has been a long tradition of trying to analyze causation in terms of counterfactuals. Such accounts run into problems with cases of causal preemption, of the sort we saw in section I above. Many ingenious proposals have been offered to deal with these problems, but the details need not concern us here. It seems clear that causation and counterfactuals are closely related,

and that in most ordinary cases, counterfactual dependence provides a pretty good *prima facie* criterion for causation. Counterfactual dependence is, for example, used as a test for causation in the law; and psychologists such as Kahneman and Tversky[21] have argued that we rely on counterfactual reasoning when making causal judgments. Thus it would hardly be surprising if the same factors that influence counterfactual reasoning also affect our causal attributions.

Our proposal is that when subjects think about the actual cause(s) of some event, they think about one event, or perhaps a small handful of events, that *made a difference* for the outcome, in the sense that hypothetical situations in which that event does not occur or in which it is dramatically different will be ones in which the outcome is dramatically different. When people select which events to hypothetically change, they are guided by the relevance of the corresponding counterfactual. If some abnormal event occurred, the hypothetical situation in which it does not occur, or some more normal event occurs instead, will typically be considered relevant. If this situation is one in which the outcome in question is substantially different, the abnormal event will be judged to be a cause of that outcome. By contrast, if some normal event occurs, we may never get around to considering the counterfactual situation in which some more abnormal alternative occurs instead.[22] Even if this situation is one in which the outcome is different, the normal event will not be judged a cause of the outcome. Thus, in the pen vignette, while it is true that the problem would not have occurred if the administrative assistant had not taken a pen, this counterfactual is not deemed as relevant as the one in which Professor Smith does not take a pen. This explains why we are less inclined to judge that the administrative assistant's action was a cause of the problem.

Note that normality and abnormality are, at least to some extent, comparative notions. Thus, even if the administrative assistants and the faculty members only infrequently take the pens, still, the administrative assistant's taking the pen is *more* normal than Professor Smith's doing so, and we are not likely to consider the possibility in which she does not take the pen.[23]

[21] Kahneman and Tversky, *op. cit.*

[22] Hart and Honoré similarly propose that judgments about the normality of a condition affect our causal judgments (see especially pp. 33–41), although they do not make the connection with the psychology of counterfactual reasoning.

[23] We may make an analogy here with Lewis's account of "Causation as Influence." In that account, it is not the absolute amount of influence that one event has over another that determines whether it is a cause, but the amount of influence it has in comparison with other candidate causes.

In the egalitarian tradition illustrated by the passages from Mill, Lewis, and Hall above, both the administrative assistant's action and Professor Smith's action would count as causes of the problem. Moreover, they are each full causes: causes do not come in degrees. But the ordinary causal judgments of subjects—even, as we shall see, those of trained philosophers—are not like this. Subjects are happy to rate their level of agreement with the causal claims using nonextreme values of a numerical scale. And to some extent at least, they tend to see the two causal claims as being in competition with one another. Even when subjects were not asked to explicitly compare the two causal claims, but only to rate their agreement with the claim that the administrative assistant caused the problem, the presence of a better causal candidate led them to disagree with this claim.

VIII. ALTERNATIVE EXPLANATIONS

Before going any further, perhaps it will be helpful to sum up the path of the argument thus far. We began by asking what purpose people's concept of actual causation might serve. To address that question, we engaged in an empirical investigation into the criteria underlying people's ordinary application of this concept. There, we argued that people's criteria for causal selection arise in part because of the way in which their judgments of counterfactual relevance are affected by *norms*. Ultimately, our aim is to figure out whether the criteria revealed by this empirical inquiry actually do serve a legitimate purpose.

But first we need to consider a certain sort of alternative explanation. On this alternative explanation, judgments about norm violation do not actually play any role in people's underlying competence, but people's causal judgments can sometimes be distorted by their judgments of *blame*. It is possible, for instance, that subjects are conflating the questions of causation and blame. Or perhaps subjects recognize that while it is strictly true that the administrative assistant caused the problem, it would be infelicitous to say so, since the context suggests that the causal inquiry is being made for the purpose of assigning blame.[24] Or perhaps we have an immediate negative affective reaction to Professor Smith's action, which biases us in favor of causal assessments that will underwrite attributions of blame to him.[25]

Although these alternative explanations can accommodate all of the data we have discussed thus far, they rely on a type of moral judgment

[24] This view is proposed by Julia Driver, "Attributions of Causation and Moral Responsibility," in Sinnott-Armstrong, ed.

[25] As proposed by Mark Alicke, "Culpable Causation," *Journal of Personality and Social Psychology*, LXIII (1992): 368–78.

that plays no role at all in our preferred account. In the pen vignette, for example, one finds two different events about which people could be making moral judgments:

Candidate cause: Professor Smith taking a pen

Effect: The problem arising

The alternative explanations suggest that people's causal judgments are influenced by a moral judgment they make about the effect—namely, a judgment that the professor is *to blame* for the problem. By contrast, our own account makes no mention of any sort of moral judgment regarding the effect. Instead, it posits a role for judgments about whether the candidate cause was itself a norm violation.

These various approaches all yield exactly the same prediction in the pen vignette, but it should be possible to develop vignettes in which the different approaches yield very different predictions. What we need now are cases in which a norm is violated but no one is assigned blame. In cases of this latter type, the alternative explanations suggest that moral considerations should have no impact on people's causal judgments (because of the absence of blame) while our own hypothesis suggests that the impact of normative considerations should remain unchanged (because people still see that a norm has been violated).

To decide between these competing views, we conducted a questionnaire study. Overall, the study had 3,422 subjects, including 327 who indicated that they were philosophy professors or had Ph.D.s in philosophy. Each subject received eleven vignettes, but only two of those vignettes were relevant to the issue under discussion here.[26]

One vignette—the *drug vignette*—concerned an agent who violates a norm but ends up bringing about an outcome that is actually *good*:

> An intern is taking care of a patient in a hospital. The intern notices that the patient is having some kidney problems. Recently, the intern read a series of studies about a new drug that can alleviate problems like this one, and he decides to administer the drug in this case.
>
> Before the intern can administer the drug, he needs to get the signature of the pharmacist (to confirm that the hospital has enough in stock) and the signature of the attending doctor (to confirm that the drug is appropriate for this patient). So he sends off requests to both the pharmacist and the attending doctor.

[26] For a full description of the study procedures and the additional vignettes, see Hitchcock and Knobe, "The Influence of Norms and Intentions on Causal Attributions," unpublished manuscript.

> The pharmacist receives the request, checks to see that they have enough in stock, and immediately signs off.
>
> The attending doctor receives the request at the same time and immediately realizes that there are strong reasons to refuse. Although some studies show that the drug can help people with kidney problems, there are also a number of studies showing that the drug can have very dangerous side effects. For this reason, the hospital has a policy forbidding the use of this drug for kidney problems. Despite this policy, the doctor decides to sign off.
>
> Since both signatures were received, the patient is administered the drug. As it happens, the patient immediately recovers, and the drug has no adverse effects.

After reading this vignette, subjects were asked to rate their agreement or disagreement with one of the following two statements:

- The pharmacist's decision caused the patient's recovery.
- The attending doctor's decision caused the patient's recovery.

Here the attending doctor is clearly violating a norm, but there is no sense at all in which he can be considered *blameworthy* for the patient's recovery. (After all, the patient's recovery is something good, and there can therefore be no question about who is to blame for it.) Yet despite the lack of blame, people were more inclined to say that the doctor's decision caused the recovery than they were to say that the pharmacist's decision caused the recovery.[27] This result indicates that people's causal intuitions can be affected by norm violations, even in the absence of any judgment of blameworthiness.

A second vignette—the *machine vignette*—described a case in which a norm is violated but there are no agents at all:

> A machine is set up in such a way that it will short circuit if both the black wire and the red wire touch the battery at the same time. The machine will not short circuit if just one of these wires touches the battery. The black wire is designated as the one that is supposed to touch the battery, while the red wire is supposed to remain in some other part of the machine.
>
> One day, the black wire and the red wire both end up touching the battery at the same time. There is a short circuit.

[27] Subjects rated each statement on a scale from 1 ('disagree') to 7 ('agree'). Overall, the statement about the attending doctor received a mean rating of 3.9; the statement about the pharmacist received a mean rating of 2.5. This difference was statistically significant, $t(3411) = 21.5$, $p < .001$. When one looks only at subjects who are philosophy professors or have Ph.D.s in philosophy, one finds the same basic pattern (mean for the attending doctor = 5.3, mean for the pharmacist = 3.5, $t(325) = 3.2$, $p < .005$).

After reading the vignette, subjects were asked to rate their agreement or disagreement with one of the following statements:

- The fact that the red wire touched the battery caused the machine to short circuit.
- The fact that the black wire touched the battery caused the machine to short circuit.

Here again, we have a case of norm violation without blameworthiness. Specifically, it seems that the red wire is violating a certain kind of norm, but it also seems clear that no one could literally *blame* the red wire for the short circuit. Nonetheless, people were more willing to say that the red wire touching the battery caused the short circuit than they were to say that the black wire touching the battery caused the short circuit.[28] So once again, it seems that people's causal intuitions are being affected by norm violations even in the absence of any judgment of blame. In addition, the fact that trained philosophers make essentially the same judgments as untutored subjects makes it much less plausible that the effect is the result of conflating causal judgments with some other kind of judgment.

IX. EVALUATING THE CRITERIA FOR CAUSATION

We have been concerned in these past few sections with purely descriptive questions about the criteria governing people's ordinary causal judgments. We began with the observation that these criteria seem to be constructed in such a way that events that involve norm violations are especially likely to be selected as causes. We then argued that this effect was best understood, not in terms of people's judgments of blame, but rather in terms of their judgments about the relevance of particular counterfactuals.

Now that we have in place a specific hypothesis about the criteria underlying people's ordinary causal intuitions, we can begin to ask whether these criteria actually serve any useful purpose. Do these criteria offer us anything beyond what we could get from causal structure alone? Is there anything helpful about selectively picking out those factors that most violate certain norms and referring to them in particular as "causes"?

[28] Subjects rated each statement on a scale from 1 ('disagree') to 7 ('agree'). Overall, the statement about the red wire received a mean rating of 4.9; the statement about the black wire received a mean rating of 2.7. This difference was statistically significant, $t(3410) = 30.2$, $p < .001$. When one looks only at subjects who are philosophy professors or have Ph.D.s in philosophy, one finds the same basic pattern (mean for the red wire = 5.3, mean for the black wire = 3.5, $t(325) = 8.4$, $p < .001$).

When one first hears that norm violations are more likely to be picked out as causes, it is natural to assume that the purpose of this mechanism must have something to do with picking out the agents who are truly *to blame* for an outcome. We think that this view is not quite right, for reasons described in the previous section, but we do believe that it is headed in more or less the right direction. One can regard the act of blaming a person as one way of intervening on that person's behavior and trying to get him or her to change. Indeed, Schlick attempted to understand all of our moral attributions in these broadly utilitarian terms.[29] While Schlick's account is widely believed to have been discredited,[30] it certainly seems that one of the things we are doing when we blame someone is to identify a person whose behavior is in need of corrective intervention. What we want to suggest, however, is that people's criteria for actual causation are not best understood in terms of this one very specific form of intervention. Instead, these criteria are best understood in terms of the broader question as to which strategies of intervention are most worth considering. So the question here is not just "Which agent in this situation is to blame?" but rather something like: "Of all the various ways in which one could have prevented this from happening, which one is the best one to focus on?"

In the kinds of cases we have been discussing, an outcome has arisen, and it is immediately clear that there are a number of different ways in which people could have intervened to prevent it. One obvious strategy in these cases would be to intervene on an abnormal aspect of the situation and make it more normal. But there is also another possibility. One could leave all of the abnormal aspects of the situation in place but then intervene to introduce additional abnormalities that prevent the outcome from occurring. As far as causal structure is concerned, these two possibilities are entirely symmetric—either type of intervention would have successfully prevented the effect. And yet, interestingly, people's ordinary causal intuitions are not perfectly symmetric. People tend to specifically pick out the abnormal factor as a cause, thereby directing attention to the possibility of intervening on the abnormal factor rather than on any of the normal ones.

[29] Moritz Schlick, *The Problems of Ethics*, translated by D. Rynin (New York: Prentice Hall, 1939).

[30] See, for example, P.F. Strawson, "Freedom and Resentment," *Proceedings of the British Academy*, XLVIII (1962): 1–25; but see also M. Vargas, "Moral Influence, Moral Responsibility" in Nick Trakakis and Daniel Cohen, eds., *Essays on Free Will and Moral Responsibility* (Newcastle, UK: Cambridge Scholars Press, forthcoming), for a reply to Strawson.

What we want to show now is that this approach to causal selection really does make sense.[31] There truly are good reasons to intervene on the abnormal factors rather than on the normal ones. Indeed, we see at least three distinct reasons here, corresponding to different types of norm.

1. Statistical Norms. Suppose that an event has occurred and we are thinking about what could have been done to prevent it. In such a case, we will presumably not be satisfied just to find an intervention that happens to work in this one specific situation. What we will want is a strategy of intervention that is *generalizable*—a strategy that would be effective not just in this one situation but also in other situations of a roughly similar type.

The criteria underlying people's judgments of actual causation appear to be designed in a way that enables them to achieve this objective. These criteria direct our attention away from interventions that work by leaving in place some highly unusual aspect of the situation and then capitalizing on it to achieve a particular effect. Instead, they direct our attention toward interventions that work by leaving in place the most normal aspects of the situation and then changing the abnormal aspects in a way that makes them more normal.

To illustrate the key issues here, we can return to our example of the reviewer who rejects a paper on the grounds that one should never use the word 'and' more than three times per page. Now consider two different strategies one could use if one wanted to intervene to prevent this paper from being rejected:

(a) making sure the paper is sent to a reviewer who does not have these bizarre views
(b) allowing the paper to be sent to a reviewer with these bizarre views but then compensating by introducing an additional abnormality, namely, making sure that the paper itself contained a highly unusual absence of uses of the word 'and'.

From the perspective of causal structure, these two strategies of intervention are perfectly symmetric—either intervention would be effective in preventing the effect. Yet there is clearly a marked difference in generalizability. If our goal is to prevent papers from getting rejected, it would *generally* be a good strategy to make sure they are not sent to

[31] Note that our concern throughout this section is with evaluative questions about whether the criteria serve a legitimate purpose, not with empirical questions about how those criteria actually arose. Nothing in what we say here should be taken as an endorsement of the adaptationist claim that the criteria underlying people's causal judgments actually arose *because* they serve a legitimate purpose.

reviewers with bizarre views, while it certainly would not generally be a good strategy to make sure that they contain fewer than three uses of the word 'and' per page. This difference trades on the abnormality of the reviewer's evaluation criteria in the statistical sense of 'abnormal'. Since one will only infrequently encounter reviewers with such idiosyncratic criteria, it does not make sense to tailor one's papers to fit those criteria. As we noted above, however, prescriptive norms often serve as heuristics when frequency information is not readily available; to that extent, it also makes sense to focus on interventions that will work in situations that are normal in the prescriptive sense as well.

The thing to notice now is that the criteria underlying ordinary judgments of causation are designed in such a way that they direct our attention to strategies of intervention that are generalizable in this way. Thus, if someone asked us to explain why the paper was rejected, we would not tend to consider counterfactuals that involved keeping in place all the abnormal factors and then adding additional abnormalities to counteract them. Instead, we would immediately think of counterfactuals that involved preventing the outcome by changing the abnormal aspects of the situation in a way that made them more normal, and we would therefore answer: "The cause of the problem here was that the paper was sent to a reviewer with bizarre views."

Interestingly, the same line of reasoning will also lead us to focus on abnormal features of the situation in the case where the outcome is good, and to be promoted. Consider the drug vignette. Suppose that as a result of this episode, the hospital administration wanted to encourage the use of the drug to treat kidney problems. Whose behavior would they need to encourage? Encouraging the pharmacists to sign off on requests to use the drug would have little effect. For one thing, the pharmacists are likely to be signing off on such requests already. Furthermore, their doing so is unlikely to result in patients being given the drug, since the attending physicians will not, as a rule, sign off on such requests. However, if the hospital administrators encourage the attending physicians to sign off on requests to use the drug, this will almost certainly result in more patients receiving the drug. So once again it makes sense to target the abnormal condition for intervention. This is what we do when we identify the attending physician's decision as a cause of recovery.

2. Moral Norms. Thus far, we have been considering the importance of picking an intervention that will generally prove effective, but there are also aspects of choosing the best intervention that are completely independent of this issue of effectiveness. Suppose, for example, that one comes upon a situation that has certain very good aspects and certain very bad ones. It might turn out that one can pre-

vent a particular effect either by making one of the bad aspects better or by making one of the good aspects worse. In such a case, there is a clear sense in which it would be better to intervene by changing the bad aspects than by changing the good ones. Notice, however, that the superiority of this type of intervention has nothing to do with its probability of successfully preventing a given effect. Even if two strategies of intervention are precisely equal in the degree to which they would successfully prevent a certain type of effect in future cases, we might still prefer the strategy that involves changing bad aspects of a situation in ways that make them better.

The striking thing now is that the criteria underlying people's ordinary causal judgments are designed in such a way that they tend to pick out precisely these sorts of interventions. Experimental results suggest that people's causal judgments truly can be influenced by their judgments about goodness and badness. Thus, one recent experiment showed that people who are pro-life and those who are pro-choice actually arrive at different causal judgments regarding acts intended to bring about the death of a fetus.[32] In a case like this one, it seems clear that people's intuitions about causation are being affected by full-blown judgments about goodness and badness.

Something like this effect may also be at work in the case of some policy norms. Consider the pen vignette once again. There, it would be possible to prevent the problem using either of two strategies:

(a) stopping the professor from doing things he should not do
(b) allowing the professor to keep doing those things but then inconveniencing the administrative assistant in a way that counteracts the professor's impact on the situation.

Although both of these strategies of intervention would be perfectly effective, there are clear reasons to prefer strategy (a) over strategy (b). The policy forbidding faculty members from taking the pens is sufficiently reasonable that we judge that it is wrong for Professor Smith to take the pen. And, as we saw above, people's intuitive judgments specifically pick out the professor as a cause, thereby directing attention toward the strategy of intervention that truly would be preferable.

3. Norms of Proper Functioning. The considerations raised in the previous section apply when the violation of a prescriptive norms is actually "bad." This will not be the case in all of the examples we introduced above. While it seems reasonable to suggest that the profes-

[32] F. Cushman, Knobe, and Sinnott-Armstrong, "Moral Appraisals Affect Doing/Allowing Judgments," *Cognition*, CVIII (2008): 281–89.

sor who unfairly takes a pen is doing something bad, when it comes to cases like an incorrectly placed wire in the machine vignette, it seems odd to use this sort of characterization. The position of the wire is not truly *bad* in any deep sense. The norm that it violates is a norm of proper functioning, rather than a moral norm. Some policy norms within social institutions might be more naturally assimilated to these kinds of norms, rather than moral norms. For example, suppose that a company manufactures very expensive dolls that most families cannot afford. The owners of the company are very wealthy and corrupt. The employees within this company are governed by rules that are designed for the efficient production of these dolls. We might think that it would not be morally wrong for an employee to violate these rules: no innocent people will be substantially harmed by the resulting inefficiencies. So the company's policies do not have the status of moral norms. However, since the policies are conducive to the efficient operation of the company, they play much the same role that norms of proper functioning do in machines and organisms. The machine vignette shows that these kinds of norms can also influence people's causal judgments. What we need to show, then, is that these sorts of judgments also play a helpful role in selecting strategies of intervention.

To illustrate the key issues here, we can introduce another example. Imagine a factory that is governed by a rule according to which Lauren is supposed to put oil in the machines while Jane is not. Now suppose that Lauren forgets to put oil in the machines and it therefore ends up breaking down. If we wanted to intervene to prevent the breakdown, we could adopt either of two possible strategies:

(a) intervene to make Lauren do what she is supposed to do
(b) allow Lauren to avoid doing what she is supposed to but then intervene to make Jane do the job instead.

Either of these interventions would be effective in preventing the machine from breaking down, but is there some reason for choosing one over the other? We believe that there is. Regardless of whether you think that there is anything bad about Lauren not doing the job she is supposed to do, it seems that there are good reasons to prefer an intervention that does not introduce additional abnormalities into the system. Given that the whole production line was set up on the assumption that Jane would not be adding oil to the machines, we have reason to worry that something else will go wrong if we intervene to make her do that job. Perhaps the problem will be that she is no longer able to work at her assigned task, or perhaps it will be that there are now too many people in the room with the machines, or perhaps it will be something else entirely. Moreover, consistently enforcing one partic-

ular set of norms is a very efficient way of achieving the end of a well-functioning factory. If Jane and Lauren each do their jobs, it is not necessary for them to coordinate their actions, or to have a deep understanding of the machine's functioning. By contrast, if we allow Lauren put oil in the machine or not, Jane will not know whether to put oil in the machine without coordinating with Lauren, or inspecting the machine to determine whether it needs oil (something she may not know how to do). It definitely does seem that we have good reason to prefer the intervention that restores the workings of the factory to precisely the way they were designed.

And, just as one might expect, subjects who are given the story of Lauren and Jane agree with the claim that "Lauren caused the machine to break down" but disagree with the claim that "Jane caused the machine to break down."[33] Here as elsewhere, it seems that people's causal intuitions are directing their attention toward the best strategy of intervention.

A similar approach can be used to explain subjects' judgments in the machine vignette. The significance of the norm violation here is not that the red wire is in any way to blame but rather that it is the most suitable target of intervention. The short circuit could be avoided by leaving the red wire in place and moving the black wire to some other part of the machine. But this might introduce new problems. Perhaps the black wire does not have the same current capacity as the red wire. In order to restore the machine to working order by intervening on the black wire, one would have to know more than just the design of the machine; one would need to understand the basic principles of its operation. Moreover, if different electricians are working on different parts of the machine, they would need to coordinate in order to make sure that only one of the wires made contact with the battery. None of this is necessary, however, if the electricians are simply instructed to assemble the machine according to its original design.

Looking through these three separate points, one may have a certain inclination to complain: "These points just don't have anything to do with each other—the argument here is nothing but a list of three unrelated claims." But one may also have an inclination to make precisely the opposite sort of complaint: "These points are just repeating each other—saying the same sorts of thing in three slightly different ways."

[33] Knobe, "Cognitive Processes Shaped by the Impulse to Blame," *Brooklyn Law Review*, LXXI (2005): 929–37. Thirty-five subjects were given each version of the vignette, and then asked to rank their level of agreement with the causal claim on a scale from zero to six. The mean rating for Lauren was 3.34, for Jane 0.37 ($p < .001$).

Perhaps each of these views contains part of the truth. The sense in which our three points are very different is that they each appeal to a different type of norm—the first to purely statistical norms, the second to norms of goodness and badness, the third to norms of proper functioning. The sense in which they are all more or less the same is that they all show the same basic structure. In each case, we have argued that it is preferable to choose the intervention that works by targeting abnormal aspects of the situation and replacing them with more normal ones. It is this basic structure, we believe, that gives our criteria for causal selection their distinctive purpose and value.

X. CONCLUDING SUMMARY

We began with the question as to whether the concept of actual causation could be shown to have any legitimate purpose. In hopes of addressing this question, we proceeded in two steps.

First, we engaged in a purely descriptive attempt to understand the criteria governing people's ordinary causal intuitions. There, we suggested that people's causal intuitions are determined in part by judgments about the relevance of counterfactuals and that judgments of relevance are, in turn, determined in part by the application of norms.

Second, we asked whether the criteria uncovered in this first part of our inquiry might serve any legitimate purpose. There, we argued that these criteria do serve a purpose—namely, that they are designed in such a way that they tend to direct us toward the best strategies of intervention.

It should be evident by now that, even if people do a perfect job of following their criteria for causal selection, they will not always end up picking out the absolutely optimal strategy of intervention. That, however, is not the point. The point is that the criteria are designed in such a way that they generally tend to direct us toward strategies of intervention that would be preferable—and this, we believe, is ample justification for our concept of actual causation.

CHRISTOPHER HITCHCOCK

California Institute of Technology

JOSHUA KNOBE

Yale University