Portable Causal Dependence: A Tale of Consilience

Christopher Hitchcock*†

This article describes research pursued by members of the McDonnell Collaborative on Causal Learning. A number of members independently converged on a similar idea: one of the central functions served by claims of actual causation is to highlight patterns of dependence that are highly portable into novel contexts. I describe in detail how this idea emerged in my own work and also in that of the psychologist Tania Lombrozo. In addition, I use the occasion to reflect on the nature of interdisciplinary collaboration in general and on the interaction between philosophy and psychology in particular.

The McDonnell Collaborative on Causal Learning is a research group spearheaded by Alison Gopnik, a developmental psychologist from Berkeley. It is funded by a multiyear grant from the James S. McDonnell foundation and has the aim of studying causal learning, reasoning, inference, and judgment. The collaborative consists of a dozen faculty members from across the United States in both philosophy and psychology, together with a number of graduate students and postdocs. Several of these—Gopnik, David Danks (philosophy, Carnegie-Mellon), Tania Lombrozo (psychology, Berkeley), and James Woodward (history and philosophy of science, Pittsburgh)—will figure in my story. The members of the collaborative met several times a year for workshops, as well as events that piggybacked on other conferences (including PSA 2010 in Montreal). In this short article, I want to describe one idea that emerged from the collaboration, situate some of my own research within this context, and reflect on the nature of interdisciplinary collaboration in general and on the interaction between philosophy and psychology in particular.

*To contact the author, please write to: Division of Humanities and Social Sciences, 101-40 California Institute of Technology, Pasadena, CA 91125; e-mail: cricky@caltech.edu.

†For comments and suggestions, I would like to thank David Danks, Alison Gopnik, Joseph Halpern, Joshua Knobe, Tania Lombrozo, Joshua Tenenbaum, and James Woodward.

Philosophy of Science, 79 (December 2012) pp. 942–951. 0031-8248/2012/7905-0015$10.00
Copyright 2012 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/667899 Published online by Cambridge University Press

There are many different models of interdisciplinary collaboration. One model involves members of different disciplines coauthoring a paper (or multiple papers), with a strict division of labor along disciplinary lines. A slightly different model is described by Galison (1998), in which members of different disciplines exchange knowledge and expertise in a form of trade. Galison describes the creation of 'trading zones' and 'pidgin' languages in analogy with commerce between different cultures.

The example that I will describe does not fit either of these models. Rather, it involves researchers from different fields pursuing independent lines of research, who nonetheless converge on a similar idea, due in part to mutual influence over an extended series of interactions and exchanges. The many workshops involving the project's participants often had the character of 'brainstorming' sessions. A number of features made these successful. The workshops were designed for presenting work in progress—where feedback is especially helpful—rather than finished work. The same participants (with minor variations) returned to the workshops over the course of several years, allowing them to establish a rapport with one another. While participants came from different disciplines, the project was focused around a set of questions and problems that the participants had a joint interest in solving. The model was thus close to Neurath's conception of 'unity of science at the point of action' (Cat, Chang, and Cartwright 1996), where the complexity of a phenomenon forces practitioners of different disciplines to cooperate and communicate successfully in order to solve scientific problems.

Successful collaboration required abandoning the territoriality that is common in philosophy. While there may be distinctively philosophical questions that will not yield to empirical data, there are also questions that lie at the intersection of philosophy and psychology. I will give a specific illustration below. Collaboration requires allowing outsiders to lay claim to a problem and bring to bear their own tools. Sometimes this requires a process of productive reinterpretation. Philosophical ideas are rarely neatly packaged in a form that is of direct help to psychology and vice versa. In discussion, often an idea presented (e.g.) by a philosopher would be reinterpreted by a psychologist by looking at the idea through a different lens and reconceiving it in terms of the concepts and needs of psychology. Sometimes this reinterpretation was deliberate, sometimes not. When our students do this, we attempt to correct them. But the product of such a reinterpretation was often interesting and useful. Allowing this to happen required giving a kind of permission for outsiders to appropriate philosophical ideas and reshape them for their own purposes.

To understand the question that forms the core of my illustration, let us begin with the familiar philosophical problem of trying to analyze actual causation. This is the relation that is described by statements, typically ex-

https://doi.org/10.1086/667899 Published online by Cambridge University Press

pressed in the past tense, claiming that one particular event caused, or was a cause of, a second event. For example, 'Suzy's throwing her rock at the window caused it to shatter'. Actual causation has sometimes been called 'singular causation' or 'token causation'; it has also been called 'causation' simpliciter. For reasons we shall see, I think these expressions are all misleading.

A familiar problem for counterfactual theories of actual causation (and other approaches based on the notion of difference making) is that of how to handle cases of preemption. Suppose that Suzy throws her rock at a window, and it breaks. Billy is standing by, rock in hand, willing and able to throw if Suzy does not. If Suzy had not thrown her rock, Billy would have thrown, and the window would have shattered just the same. Suzy's throw causes the window to shatter, but the window would have shattered even if Suzy had not thrown.

We can represent the structure of this case simply using a structural equation model. Let SR take the value 1 if Suzy is ready to throw, 0 if not, and analogously for BR. Let ST take the value 1 if Suzy throws, 0 if not, and analogously for BT. Finally, let W take the value 1 if the window breaks, 0 if not. Then, our equations will be

$$\mathrm{SR} = 1;$$

$$\mathrm{BR} = 1;$$

$$\mathrm{ST} = \mathrm{SR};$$

$$\mathrm{BT} = \mathrm{BR} \times (1 - \mathrm{ST});$$

$$W = \max(\mathrm{ST}, \mathrm{BT}).$$

Each equation describes how a variable depends on its immediate causes. We represent a counterfactual by replacing the equation for the appropriate variable. For example, to evaluate what would have happened if Suzy had not thrown, we replace the third equation with $\mathrm{ST} = 0$. This gives us $\mathrm{SR} = 1$, $\mathrm{BR} = 1$, $\mathrm{ST} = 0$, $\mathrm{BT} = 1$, and $W = 1$. Note that we do not substitute 0 for ST in the third equation; this would lead to $\mathrm{SR} = 0$. This way of representing counterfactual suppositions ensures that counterfactuals do not 'backtrack' (in the sense of Lewis 1979).

Other representations are possible. In Hitchcock (2007b) I give my reasons for preferring structural equation models to the commonly used 'neuron diagrams', but the latter would also serve my current purpose.

There is an intuitive sense in which the structural equation model represents the causal structure of the example. In particular, the equations represent asymmetric relations of dependence and not merely regularities or correlations. But one cannot simply read off from the model that Suzy's throw is

https://doi.org/10.1086/667899 Published online by Cambridge University Press

an actual cause of the window's shattering. In effect, then, the traditional philosophical problem (as engaged, e.g., by David Lewis 1973) is how to define actual causation in terms of causal structure.

Early on in our discussions, a question emerged: What function is served by the concept of actual causation? The structural equation model is a simple and elegant model of the causal structure of the case. We can use the model to predict what would occur, given any starting conditions. We can use the model to predict the effect of any interventions one might perform (e.g., preventing Suzy from throwing). And one can use the model to evaluate the truth of any counterfactual supposition. What more could we want from a representation of the causal relations that hold in a particular example? And yet, in addition to the model, we also make the judgment that Suzy's throw caused the window to shatter. Why do we apply this further concept, rather than just making do with the model?

I think it is interesting to examine the preconditions for asking this kind of question, as well as the different kinds of disciplinary concerns that make it pressing. The question is made possible, in part, by the availability of a representation that seems to give us most of what we want from causal knowledge, without explicitly encoding information about actual causation. Only against the background of such a representation does the concept of actual causation appear to stand out as a superfluous addition.

This kind of question certainly has a philosophical pedigree. For example, Carnap's (1950) project of explication involved constructing concepts that were well designed for specific tasks. Understanding what work the concept of actual causation is supposed to do would help to provide a normative framework for evaluating potential explications of the concept. For another example, the sort of account of objective chance offered by Lewis (1980) might best be understood as an account of the role that beliefs about chances play in our cognitive economy.

The question is also motivated by the concerns of the cognitive scientist for cognitive efficiency. A representation of actual causation would command cognitive resources, such as attention and processing time, and presumably also physical resources, such as blood, oxygen, and calories. Why pay these prices with no corresponding gain?

While no conscious effort was made to arrive at a consensus answer to this question, a number of members of the collaborative coming to this problem from different directions (and in some cases, working with collaborators from outside the McDonnell Collaborative) arrived at very similar answers. The basic idea is that judgments of actual causation pick out patterns of dependence that are highly 'portable' into contexts beyond that of the original case. Such patterns of dependence are more likely to be encountered again and may perhaps even be exploited to produce desired outcomes in novel situations. This conclusion is counterintuitive. One of the characteristic fea-

https://doi.org/10.1086/667899 Published online by Cambridge University Press

tures of actual causation is supposed to be that it is a relation between particular events. As mentioned above, actual causation has sometimes been called 'singular causation' or 'token causation'. It is often explicitly contrasted with 'general' or 'type-level' causation, as described in claims such as 'throwing rocks at windows causes them to shatter'. This is one reason that I think the label 'singular causation' is inappropriate. (Another is that the causal structure described by the structural equation model is also singular: its variables represent Suzy and Billy, on this particular occasion, throwing rocks [or not] at this particular window.)

Important antecedents of this kind of view can be found, for instance, in Gopnik (2000), Lombrozo and Carey (2006), and Woodward (2006; all except for Carey are members of the McDonnell Collaborative). While these papers all discussed the portability or stability of causal and explanatory relations, they were not specifically theses about actual causation. More mature formulations of the kind of view I have in mind are found in Hitchcock and Knobe (2009), Lombrozo (2010), Halpern and Hitchcock (2012), and Rose, Danks, and Machery (2012). I will here describe Tania Lombrozo's work as well as my own.

Lombrozo came to these issues with an interest in different kinds of explanation, in particular, teleological explanations versus causal/mechanical explanations. She was interested in the conditions under which people accepted teleological explanations. Lombrozo and Carey (2006) argued that the acceptability of a teleological explanation was, partly, a function of its portability.

In Lombrozo (2010), she considered two related questions: (i) Under what conditions are intentional actions more likely to be judged causes than unintentional actions? (ii) Under what conditions are events connected to an outcome by a physical process (call them 'process connectors') more likely to be judged causes than double preventers (which prevent something that would have prevented the outcome; see Hall 2004)? The latter comparison drives a wedge between process theories of causation (e.g., Salmon 1984; Dowe 2000) and counterfactual or difference-making conceptions of causation. The former claim that process connectors are causes, while double preventers are not, while the latter claim that both are causes. For example, suppose A throws a ball at the 'on' switch of a radio (a direct physical process), while C steps on the power cord so that B cannot unplug it (a double preventer). When are subjects more likely to judge that A is a cause of the resulting music than C? What effect does it have if A or C is acting intentionally or unintentionally. (A was juggling, which would explain the unintentional throwing of the ball.) In her experiment 1a, Lombrozo presented subjects with this example or with another that had a similar structure. She varied whether A or C (or both or neither) were acting intentionally. She then asked subjects to rate, on a six-point scale (1 = totally inappropriate;

https://doi.org/10.1086/667899 Published online by Cambridge University Press

6 = totally appropriate), the appropriateness of saying that A or C caused the music to play.

Lombrozo found that there was an interaction effect between the two variables. When the agents acted intentionally, it made no significant difference whether they threw the ball at the radio or prevented unplugging by stepping on the power cord. And when the agents threw the ball at the radio, it made no significant difference whether they did so intentionally. However, when the agents stepped on the power cord, it made a significant difference whether they did so intentionally, and when agents acted unintentionally, it made a significant difference whether they were process connectors or double preventers.

Lombrozo explained this effect in terms of the portability of the dependence relations that obtain in each case. Patterns of dependence that are mediated by direct connecting processes are portable because they are insensitive (in the sense of Woodward 2006). In order to turn on a radio by throwing a ball at the on switch, the radio has to be plugged in, but not much else is needed. A's throw would have worked even if B and C had been absent, if they had been doing something else, and so on. By contrast, stepping on the power cord will be a means of turning on a radio only in fairly specific circumstances: it requires that someone or something press the on switch and that someone or something else threaten to unplug it. If either A or B or both had been absent, C's action would have been ineffective: either the radio would not have come on (if A were absent), or else it would have come on regardless of C's action (if A were present but B absent). Thus, the dependence of the outcome on C's action was highly sensitive.

However, when an outcome is brought about intentionally, the dependence of the outcome on the intention is portable because intentions are flexible. If the peculiar circumstances had not occurred, C could have found some other means for turning on the radio. It is only in the case in which C unintentionally stepped on the power cord that we have a dependence relation that is not readily portable to other situations.

I was led to a different version of the portability view through an interest in the role of normality in causal judgments. 'Normality' is deliberately ambiguous, having connotations of 'frequency' or 'typicality' but also of conformity with prescriptive norms of various sorts: morality norms, laws and policies, and norms of proper functioning. While these are obviously different in important ways, they seem to influence causal judgments in similar ways. In Hitchcock and Knobe (2009), Josh Knobe and I speculated that these different conceptions of normality often serve as heuristics for one another. The idea is that an agent will typically approach a situation with a script or template describing how things 'ought' to go.

Pressure to include considerations of normality in an account of actual causation comes from several different directions. At around the same time,

https://doi.org/10.1086/667899 Published online by Cambridge University Press

Ned Hall (2007), Joseph Halpern (2008), and I (Hitchcock 2007a) all suggested that considerations of normality, or of the default behavior of a system, could be used to address some counterexamples with existing theories of actual causation. For example, consider the simple preemption example discussed earlier. While the window's breaking does not counterfactually depend on Suzy's throw in the actual world, counterfactual dependence is restored if we remove Billy (represented by setting $BR = 0$ in the model) or hold fixed the fact that Billy did not throw his rock (represented by setting $BT = 0$ in the model). However, there are other cases in which we can create counterfactual dependence by introducing changes into a model and where we do not ascribe actual causation. Hall, Halpern, and I suggested that the permissibility of such changes depends on whether they make the situation more normal or less normal. Since the presence of a backup thrower is an abnormal feature of the preemption case, it is permissible to remove him. The counterfactual dependence that is present in the more normal situation suffices for actual causation.

Another point of entry was the experimental work of Josh Knobe and his collaborators on the influence of norms on causal judgment. For example, Knobe and Fraser (2008) presented subjects with a short vignette involving a philosophy department, which had a policy that faculty were not allowed to take pens from the receptionist's desk. One day, Professor Smith and an administrative assistant take the last two remaining pens, resulting in a problem when the receptionist cannot take a message. Subjects were then asked to rate their agreement or disagreement (on a seven-point scale) with either the statement that Professor Smith caused the problem or the statement that the administrative assistant caused the problem. Subjects agreed much more strongly with the claim that Professor Smith caused the problem. In this example, Professor Smith violated the department's policy, while the administrative assistant did not, and this difference had a strong effect on causal judgments. Cushman, Knobe, and Sinnott-Armstrong (2008) obtained a similar result using moral norms, while Hitchcock and Knobe (2009) obtained the effect with norms of proper functioning. In addition, Hitchcock and Knobe (2009) showed that the effect occurs, even if the outcome is good or morally neutral.

In Hitchcock and Knobe (2009), Josh Knobe and I explained this effect by claiming that judgments of actual causation pick out factors that make a difference in normal circumstances. This is desirable because such dependence relations are highly portable. In particular, such dependence relations will figure prominently in the script of a template that an agent will carry around and use as a guide in novel cases.

In collaboration with Joseph Halpern (Halpern and Hitchcock 2012), I have proposed a 'graded' notion of actual cause (Moore [2009] also proposes a scalar notion of cause). Potential causes are graded according to the normality of the situation in which they make a difference for the effect.

https://doi.org/10.1086/667899 Published online by Cambridge University Press

Suzy's throw is considered an actual cause of the window shattering because her throw would make a difference in the normal situation in which no backup thrower is present. In Knobe and Fraser's vignette, Professor Smith's action receives a higher causal grade since his action would make a difference in the normal case in which the administrative assistant takes a pen, whereas the administrative assistant's action makes a difference only in the abnormal case in which Professor Smith takes a pen. Our account also lets us distinguish between direct causes and double preventers in much the same way that Lombrozo's does.

The data that drive these accounts are different from the intuitive responses that are usually used to defend conceptual analyses in philosophy. First, the subjects who participated in the studies were not (all) philosophers. However, Hitchcock and Knobe (2009) did collect data on over 300 subjects who identified themselves as philosophers (defined as holding a PhD in philosophy or holding an appointment as a professor of philosophy). The causal judgments of these respondents were influenced by norms in the same way that other subjects' responses were.

A second difference is that subjects in these studies were not presented with a simple yes/no question but rather given an opportunity to rate their level of agreement on a numerical scale. A related difference is that the data are heterogeneous: different subjects gave different responses. By contrast, when intuitions are offered in support of conceptual analyses, the normal assumption is that these intuitions are shared. Ludwig (2007), for example, argues that anyone who arrives at a judgment about the applicability of a philosophical concept in the appropriate way, relying only on her competence with the concept in question, will arrive at the same answer. The result of both of these differences is that the phenomena to be explained are significant differences in subjects' average responses, rather than sharp yes/no verdicts. This is not especially problematic if one thinks that actual causation can come in degrees, as claimed by Moore (2009) and Halpern and Hitchcock (2012; see also Hitchcock 2003). However, the usual assumption has been that one event either is or is not an actual cause of another, with no room for matters of degree.

Third, the judgments of subjects exhibit patterns that philosophers have long warned about. Mill (1843) and Lewis (1973), for example, note that people often discriminate among the various causes of some event and call one 'the cause'. But both Mill and Lewis warn that that discrimination (which Lewis calls 'invidious') is not relevant to the metaphysics of causation. So perhaps it is not surprising that considerations such as intentions and norms affect people's causal judgments, but this does not tell us anything about the nature of causation. Although neither Mill nor Lewis makes the case explicitly, it is reasonable to assume that the central reason for rejecting subjects' discrimination as invidious is that actual causation is sup-

https://doi.org/10.1086/667899 Published online by Cambridge University Press

posed to be an objective feature of the world. Moore, for example, dismisses the idea that intentions can affect the causal status of an event as a belief in 'telekinesis' (2009, 137).

Just as the existence of a simple, precise causal model played a role in motivating the question about the role of the concept of actual causation, I think that it can play a role in defusing these worries. The model describes the underlying causal structure of the case. This is a purely objective feature of a system, determined by the truth values of nonbacktracking counterfactuals (Lewis 1979) or by the results of potential interventions on the system (Woodward 2003). This causal structure is the locus of the objectivity of causation. This underlying, objective structure, however, does not include relations of actual causation. As we noted already, relations of actual causation cannot simply be read off the causal model. Rather, relations of actual causation are matters of degree that are determined in part by the objective causal structure of a system and in part by factors such as norms and intentions. The function of ascribing relations of actual causation is to locate dependence relations that are highly portable to other systems.

REFERENCES

Carnap, R. 1950. *Logical Foundations of Probability.* Chicago: University of Chicago Press.

Cat, J., H. Chang, and N. Cartwright. 1996. "Otto Neurath: Politics and the Unity of Science." In *The Disunity of Science: Boundaries, Contexts and Power*, ed. P. Galison and D. Stump. Stanford, CA: Stanford University Press.

Cushman, F., J. Knobe, and W. Sinnott-Armstrong. 2008. "Moral Appraisals Affect Doing/Allowing Judgments." *Cognition* 108:281–89.

Dowe, P. 2000. *Physical Causation*. Cambridge: Cambridge University Press.

Galison, P. 1998. *Image and Logic*. Chicago: University of Chicago Press.

Gopnik, A. 2000. "Explanation as Orgasm and the Drive for Causal Knowledge: The Function, Evolution, and Phenomenology of the Theory-Formation System." In *Explanation and Cognition*, ed. F. Keil and R. A. Wilson, 299–324. Cambridge: Cambridge University Press.

Hall, N. 2004. "Two Concepts of Causation." In *Causation and Counterfactuals*, ed. J. Collins, N. Hall, and L. Paul, 225–76. Cambridge, MA: MIT Press.

———. 2007. "Structural Equations and Causation." *Philosophical Studies* 132:109–36.

Halpern, J. 2008. "Defaults and Normality in Causal Structures." In *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference*, ed. Gerhard Brewka and Jérôme Lang, 661–72. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Halpern, J., and C. Hitchcock. 2012. "Graded Causation and Defaults." Unpublished manuscript, Department of Computer Science, Cornell University.

Hitchcock, C. 2003. "Of Humean Bondage." *British Journal for the Philosophy of Science* 54: 1–25.

———. 2007a. "Prevention, Preemption, and the Principle of Sufficient Reason." *Philosophical Review* 116:495–532.

———. 2007b. "What's Wrong with Neuron Diagrams." In *Causation and Explanation*, ed. J. Campbell, M. O'Rourke, and H. S. Silverstein, 69–92. Cambridge, MA: MIT Press.

Hitchcock, C., and J. Knobe. 2009. "Cause and Norm." *Journal of Philosophy* 106:587–612.

Knobe, J., and B. Fraser. 2008. "Causal Judgment and Moral Judgment: Two Experiments." In *Moral Psychology*, vol. 2, *The Cognitive Science of Morality*, ed. W. Sinnott-Armstrong. Cambridge, MA: MIT Press.

https://doi.org/10.1086/667899 Published online by Cambridge University Press

Lewis, D. 1973. "Causation." *Journal of Philosophy* 70:556–67. Repr. with postscripts in *Philosophical Papers* (Oxford: Oxford University Press, 1986), 2:159–213.
———. 1979. "Counterfactual Dependence and Time's Arrow." *Noûs* 13:455–76. Repr. in *Philosophical Papers* (Oxford: Oxford University Press, 1986), 2:32–52.
———. 1980. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability*, vol. 2, ed. R. C. Jeffrey. Berkeley: University of California Press. Repr. with postscripts in *Philosophical Papers* (Oxford: Oxford University Press, 1986), 2:83–132.
Lombrozo, T. 2010. "Causal-Explanatory Pluralism: How Intentions, Functions, and Mechanisms Influence Causal Ascriptions." *Cognitive Psychology* 61:303–32.
Lombrozo, T., and S. Carey. 2006. "Functional Explanation and the Function of Explanation." *Cognition* 99:167–204.
Ludwig, K. 2007. "The Epistemology of Thought Experiments: First Person vs. Third Person Approaches." *Midwest Studies in Philosophy* 31:128–59.
Mill, J. S. 1843. *A System of Logic*. London: Parker.
Moore, M. 2009. *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. Oxford: Oxford University Press.
Rose, D., D. Danks, and E. Machery. 2012. "The Reasons Model of Actual Causation Judgments." Unpublished manuscript, Department of Philosophy, Rutgers University.
Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.
Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
———. 2006. "Sensitive and Insensitive Causation." *Philosophical Review* 115:1–50.

https://doi.org/10.1086/667899 Published online by Cambridge University Press