# Causation: Interactions between Philosophical Theories and Psychological Research

James Woodward*†

This article explores some ways in which philosophical theories of causation and empirical investigations into causal learning and judgment can mutually inform one another.

**1.** Recent decades have seen a great deal of fruitful work on causal learning and reasoning. Some of this work has been "theoretical" and, within this category, often "normative" in character—normative in the sense that the theories in question purport to explain how people ought to learn, reason, and judge regarding causal relationships. This work includes both "theories of causation" of a traditional philosophical sort (e.g., regularity, counterfactual, and process theories) as well as "computational theories" proposed by researchers from a range of different disciplines, including philosophy, computer science, statistics, and psychology.[1] Other work, typically but not always conducted by psychologists, has been empirical—usually experimental—and concerned with how subjects (adult humans, but also small children and nonhuman animals) learn and reason about causal relationships.

My interest in this article is in the interactions, actual and potential, between these two areas of work—the theoretical/normative and the empirical. How might one best use developments in each area to inform the other?

*To contact the author, please write to: Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA 15260; e-mail: jfw@pitt.edu.

†Thanks to Alison Gopnik and the other members of the McDonnell Collaboration on Causal Learning for many stimulating discussions.

1. Examples include the constraint-based learning algorithms devised by Spirtes, Glymour, and Scheines (2000), the hierarchical Bayesian learning models devised by Tenenbaum and his collaborators (e.g., Griffiths and Tenenbaum 2009), and Cheng's causal power theory (1997).

Philosophy of Science, 79 (December 2012) pp. 961–972. 0031-8248/2012/7905-0002$10.00

This topic is (I hope) of interest for several reasons. First and most obviously, it matters for understanding causal reasoning itself. Second, there has been an upsurge of interest about the relevance of empirical work to traditional philosophical projects, as indicated by the explosion of work in so-called experimental philosophy, or X-Phi. Recent work on causal reasoning provides a concrete example of a community of researchers pursuing projects that mix together the "philosophical/theoretical" and the "empirical." In what follows I describe some of this work and attempt to extract some morals from it.

**2.** I begin with some remarks about the role of normative theories in understanding causal reasoning. Some of the computational theories mentioned above are explicitly embedded in larger normative frameworks such as Bayesianism (e.g., Griffiths and Tenenbaum 2009). In other cases, they are accompanied by various sorts of theoretical arguments for the foundational principles on which they rely as well as arguments purporting to show that their inference procedures have normatively attractive properties such as good reliability characteristics (Spirtes, Glymour, and Scheines 2000). More traditional philosophical work (regularity theories, counterfactual theories, etc.) often has (much) less to say about learning but is still normative in the sense that it contains claims about the causal judgments people ought to make. For example, virtually all philosophical theories take it to be a condition of adequacy that no acceptable account of causation should judge that $C$ causes $E$ whenever $C$ and $E$ are correlated. In other cases, philosophical theories disagree in the causal judgments they recommend: many versions of counterfactual theories yield the judgment that counterfactual dependence involving double prevention (see below) is genuine causation, while causal process theories of the sort associated with Dowe (2000) deny this.

This leads to a natural question: How should we understand the relationship between such normative theories and empirical results about how people reason and judge? Suppose that experimental results show that many people do not regard relationships of double prevention as causal. Does this suggest that philosophical theories according to which such relations are causal are mistaken?

In exploring this question, I begin with the observation that, as an empirical matter, human causal cognition seems to be fairly successful in enabling us to cope with the world. To the extent this is so, it suggests several points. First, there must be some story about how we are able to do this (i.e., how we succeed in acquiring causal representations that are accurate enough to guide thought and action in many cases). This is something a combination of descriptive and normative theories should be able to tell us about. In particular, if we have a normative theory that tells us that we ought to reason

about causal relations in certain ways (ought to in the sense that if we so reason we will be successful in achieving some of our goals) and if we find people in fact reasoning in some good approximation to what is recommended, then these facts can form part of a potential explanation of why (and to what extent) people are successful causal reasoners. In other words, normative theories can suggest features F such that if (as an empirical matter) we find them in people's reasoning, their presence can help to explain why that reasoning is successful, to the extent that it is. In some cases these features may be such that it probably would not occur to us to investigate whether they are present in people's reasoning in the absence of an associated normative theory; in this sense, the normative theory may play a motivating or enabling role in leading researchers to do certain experiments. Conversely, if we find that people's causal cognition exhibits features G, we may ask whether there is some normative theory according to which G contributes to successful causal reasoning and learning. If so, we might consider the possibility that the reasoning in question is successful to the extent that it incorporates G.

This methodology is already employed to some considerable degree in investigations of causal learning, and I suggest it can also be fruitfully applied to the investigations of the kinds of causal judgments that people make.[2] For example, if people distinguish among different sorts of relationships in the sense that they judge that some and not others are "causal" or if they regard some as more "paradigmatically causal" than others, we may ask whether there are normative goals associated with causal thinking that these distinctions are serving. Thus, to the extent that many people regard relations of double prevention as noncausal or not paradigmatically causal, we may ask whether there is some normative goal that is served by such judgments.

Thinking about causal reasoning in this way involves thinking about it in broadly functional terms; that is, we proceed by assuming that there are aims, goals, or functions served by causal thinking, and we then see whether we use can normative theorizing to understand features of that thinking by seeing how those features contribute to the goals. Interventionists about causation (Woodward 2003) think that the acquisition of information relevant to manipulation and control is among the goals centrally associated with causal thinking, but of course there are other candidates for such goals—for example, information compression and simplification or unification of patterns of correlation.

Obviously, the inferences connecting the normative and descriptive I have been describing are defeasible: people do make systematic mistakes in causal

2. On causal learning, see, e.g., the papers in Gopnik and Schulz (2007).

inference and fail to make distinctions in judgment when they should. We cannot just read off what is normatively correct from people's actual inferences, and we cannot automatically assume that whatever is normatively correct will be reflected in people's reasoning. Even when people exhibit normatively good reasoning, some alternative normative account from the one we are entertaining may provide a better reconstruction of their reasoning. But the strategy just described is nonetheless a useful heuristic: if people exhibit patterns of causal reasoning with features that we can reconstruct within some normative framework as contributing to successful inference, this can both help to support the normative theory and increase our confidence that the features are playing some functional role.

To further develop this theme of the relation between the descriptive and the normative, consider a very suggestive analogy with visual perception.[3] The visual system is not just a set of mechanisms for producing "visual judgments"—instead, one of the most striking features of the visual system is that, although sometimes subject to visual illusions, it produces outputs that are largely reliable in the sense of enabling us to get around in the world successfully. Any adequate descriptive theory of the visual system needs to explain this fact—that is, to explain how, from the very limited information that impinges on the retina, the brain is somehow able to reach conclusions that are veridical enough, for many practical purposes, about the three-dimensional world of medium-sized objects that lies around us. These veridical judgments include the segmentation of the world into objects, with largely accurate perception of their shapes, colors, and other visually accessible properties. To provide explanations of how this is possible, normative theorizing (typically computational in form) is crucial—theorizing that is normative in the sense of specifying computations and algorithms that show how it is possible, for example, to derive accurate information about object individuation from information available to the visual system concerning shading, edges, and so on. My claim is that normative theorizing can play a similar role in connection with causal reasoning. If this is correct, normative theorizing about how we ought to learn and reason causally may not be as sharply separate from descriptive characterizations of how people in fact reason as some suppose.

**3.** The remainder of this article provides illustrations of the claims in section 2, using the interventionist theory of causation developed in Woodward (2003). The motivating idea of this theory is that causal relationships are relationships that are potentially exploitable for purposes of manipulation and control. More specifically,

3. I owe this analogy to Alison Gopnik.

(**M**) *C* causes *E* if and only if there is some intervention that changes the value of *C*, such that if that intervention were to occur, the value of *E* or the probability distribution of *E* would change.[4]

An intervention on *C* is an exogenous change in the value of *C* that is appropriately unconfounded from the point of view of inferring whether there is a causal connection from *C* to *E*. A randomized experiment in which some value of *C* is applied to a treatment group, and some alternative value of *C*, to a control group, and the resulting value of *E* across the two groups is compared is a paradigmatic example of an intervention. As an empirical matter, it is presumably true that some human actions qualify as interventions—see below—but changes not involving human action can also qualify as interventions if they have the right causal structure, as in so-called natural experiments. And of course human actions can be confounded in such a way that they do not qualify as interventions.

A crucial feature of this picture is that intervening on a variable is different from conditioning on it. Consider the structure in which *C* is a common cause of $E_1$ and $E_2$:

$$E_1 \leftarrow C \rightarrow E_2. \tag{1}$$

In structure (1), $E_1$ and $E_2$ are correlated: $\Pr(E_1/E_2) \neq \Pr(E_1)$. Suppose, however, we intervene to "set" the value of $E_2$ in (1) by means of an intervention *I*. We may think of this operation as replacing (1) with

$$E_1 \leftarrow C \quad E_2 \leftarrow I. \tag{2}$$

In (2), $\Pr(E_1/\text{set } E_2) = \Pr(E_1)$—that is, $E_1$ and $E_2$ are no longer correlated when $E_2$ is set by *I*, corresponding to the fact that $E_2$ does not cause $E_1$.

The characterization M does not claim that people can only learn about causal relationships by performing interventions. Instead, people may learn causal relationships from many sources, including passive observation, but, according to interventionism, the content of what is learned in all such cases is given by M: what one learns is what would happen were an intervention to be performed.

One way of thinking about M is as a purely normative theory. Thinking in this way, we might hold that if people do not actually judge and learn in conformity with M, so much the worse for them. But given my remarks above about the descriptive and the normative, it is also natural to ask whether M captures aspects of how (as a matter of empirical fact) sub-

4. Here *C* and *E* are variables, which can take different values. See Woodward (2003) for this and many other details that I cannot provide here.

jects think and learn. In particular, although it is true that people can learn causal relations from many sources, one would think that if interventionism is empirically plausible, people would find it particularly easy to learn from interventions. Thus, some natural empirical questions suggested by an interventionist account of causation are these: Can subjects learn causal relationships in a way that respects the normative connection between causation and intervention embodied in M? Do subjects learn about causal relationships more readily when given (or when they are able to generate) information about what follows from their own or other's interventions than when they are given other sorts of information—for example, purely covariational information not involving interventions that is nonetheless diagnostic of causal relationships? Are subjects sensitive to the normative difference between intervening and conditioning? As it happens, there is evidence that the answer to all three of these questions is yes for adult human beings and that the answer to at least the first two is yes for young children (e.g., Steyvers et al. 2003; Sloman and Lagnado 2005).

The examples just referred to are experiments "suggested" or motivated by a normative theory. In the absence of such normative theories, it is much less likely that it would have occurred to anyone to do these experiments—for example, to investigate whether people's causal inferences respect the difference between conditioning and intervening. At the same time, when we find, experimentally, that people behave and judge in a way that respects this distinction, this provides at least some support for the suggestion that they are indeed operating with a notion of causation that is connected to intervention in the way that M claims. This in turn suggests the connection between the causal judgments we make and our interest in manipulation and control is not just an arbitrary association made by normative theory but instead connects to features that are actually present in people's causal judgments.

**4.** I turn now to a more detailed example involving a sequence of empirical investigations "suggested" by a specifically interventionist account of causation. I claimed above that some human actions qualify, as an empirical matter, as interventions. Indeed, there is substantial empirical evidence that people, including young children, commonly behave as though they take their own uncoerced actions to have the characteristics of an intervention (people treat this as something like a default assumption) and use this to make the sorts of causal inferences that are warranted by M. That is, subjects have a strong tendency to infer that close-by outcomes that are correlated with their uncoerced actions are caused by them and are less willing to make this inference when they believe that their actions are coerced (e.g., Kushnir and Gopnik 2005). If coerced actions are more likely to be

confounded (an empirically plausible assumption in many cases), this is normatively appropriate behavior. This is not to say, of course, that most people (yet alone small children) possess an explicit normatively correct notion of intervention and consciously use this as a basis for causal reasoning. Rather, all that is required is that they are usually able to reliably recognize whether their action is self-generated and uncoerced (and hence presumptively an intervention) and on the basis of this engage in normatively appropriate causal inferences.

Given this connection between human action and interventions, and the role of the latter in causal inference/learning, as reflected in principles like M, we may distinguish conceptually among at least three possibilities (e.g., Woodward 2007). (1) *Egocentric causal learning*: a subject might be able to learn about causal relationships from her own interventions but not from other sources. That is, the subject behaves as if she grasps that there are regular, stable relationships between her own manipulations and various downstream effects but stops at this point, not behaving as though she recognizes that the same relationship can be present even when she herself does not act but other agents act similarly or when a similar relationship occurs in nature without the involvement of any agent at all. Behaviorally this might be manifested in, for example, a subject's failure to use relevant information from the interventions of others or from passive observation to design her own interventions.

(2) *Agent causal learning*: the subject grasps that the very same relationship she exploits in intervening also can be present when other agents act. Accordingly, she learns causal relationships both from observing the results of her own interventions and from observing the results of the interventions of others and integrates this information, recognizing, for example, that results of the interventions of others provide evidence for what would happen if she were to perform similar interventions. But the subject does not similarly learn causal relationships from "passive" observations of associations not involving interventions or at least fails to integrate these with what is learned from observing interventions.

(3) *Full (or more nearly full) causal learning*: the subject grasps that the very same relationship that she exploits in intervening also can be present when other agents intervene and can be present in nature even when no other agents are involved. Information from each of these sources is integrated into a common unified representation.

It seems uncontroversial that 1, 2, and 3 are conceptually distinct and that human adult causal learning/judgment involves 3. It is thus an interesting empirical question whether there exist subjects who are just egocentric or agent causal learners and not full causal learners.

In fact, the available evidence suggests that no nonhuman animals are full causal learners in the sense that they will perform appropriate novel interven-

tions on the basis of an observed association reflecting a causal connection between events not involving interventions.[5] Although many nonhuman animals are able to learn predictive relations on the basis of observed associations of events not involving interventions, as in classical conditioning, and are also capable of learning relationships between their own behavior and the outcomes these cause, as in instrumental conditioning, there is little or no evidence of transference back and forth between these two sets of abilities among nonhuman animals. It is thus possible, although by no means established, that the ability to integrate causal information based on relationships associated with interventions with causal information based on purely observational patterns of information into a common system of representations is a distinctively human feature of causal cognition.

But what about very young humans? Is there a stage in human development of causal cognition in which young children behave as egocentric or agent but not full causal learners? A recent experimental paper on which I was a coauthor (Bonawitz et al. 2010) reports results addressing this question. Compressing greatly (and omitting a number of other questions explored in the paper), the relevant results were as follows: we employed an experimental paradigm in which children in two different age groups (toddlers vs. preschoolers: mean age 47.2 vs. 24.4 months) observed a block slide toward a toy airplane that activated when the block touched the base of the airplane. In one condition (the so-called ghost condition), no agent was involved in the movement of the block—it appeared to move spontaneously, activating the plane on touching. In another condition (the "agent" condition), an experimenter moved the block, demonstrating that it would activate the plane on touching.

The children were then asked to make the airplane go by themselves, after being given experience with moving the block so that it was clear that they knew how to do this. A very large majority of preschoolers were able to use their previous observations of the association between the movement of the block to design an intervention of their own to activate the airplane by moving the block. By contrast, none of the toddlers were able to do this, despite the fact that we had independent evidence that they had learned (from the ghost condition) that there was a predictive association between movement of the block and plane activation. We knew that they had learned such a predictive association because the children engaged in predictive looking when exposed to the association between the movement of the block in the ghost condition and activation of the plane—seeing the movement of the block,

5. Waisman et al. (2009) looked for such an ability in dogs and failed to find it. Tomasello and Call (1997) describe a thought experiment requiring full causal learning and claim that nonhuman primates will fail at this task.

they looked to the plane as though they expected it to activate. But despite having learned this predictive association, they were unable to use it to design successful interventions.

By contrast, many of the toddlers were able to design an intervention of their own to use the block to activate the plane when they had previously observed the experimenter activate it. Although the matter needs more detailed investigation, this is prima facie evidence that there is a stage in the development of human causal cognition in which children (toddlers in this case) are able to learn to design their own interventions from observations of the interventions of other agents but not from the observation of otherwise similar associations not involving interventions. In other words, at this point the toddlers appear to be in something like an agent causal stage of learning. This is then followed by a stage in which fully causal learning is achieved, as evidenced by the behavior of the older children. Again, I emphasize that this is an experiment that was unlikely to be carried out unless one was motivated by the "philosophical/computational" idea that there is an important connection between causation and intervention and that some human actions are paradigmatic interventions.

**5.** I want now to switch gears and move to a different topic—some recent experimental work on "double prevention." This involves cases in which if *d* were to occur, it would prevent the occurrence of *e* (which would otherwise occur in the absence of *d*) and in which the occurrence of *c* prevents the occurrence of *d*, with the upshot that *e* occurs. In such cases there is overall counterfactual dependence of *e* on *c* (with the dependence in question being the nonbacktracking sort associated with causal relatedness according to both interventionist and Lewis-style [1973] counterfactual theories of causation). In Hall's (2004) example, Billy shoots down an enemy fighter (*c*) that otherwise would have shot down Suzy's bomber (*d*), enabling Suzy to successfully release her bombs (*e*). Although such counterfactual dependence is regarded by many, including Lewis, as sufficient for causation (and although in a type-causal analogue to this example, M would also judge the relation to be causal), a common reaction is that double prevention cases either are not cases of causation at all or at least lack some feature that is central to many other cases of causation. The presence of a "connecting process" or energy/momentum transfer is one natural candidate for such a feature.

This suggests several questions. First, as a matter of descriptive psychology, do people tend to regard cases of double prevention (or at least some cases of double prevention) as noncausal or at least as less paradigmatically causal than other sorts of causal claims? Second, to the extent this is the case, why do people judge in this way? If people distinguish between double prevention and (what they regard as) more paradigmatic cases of causal dependence, what normative goals might be served by their doing so? (I assume it

is not illuminating to say that they so distinguish because only the latter are "really" causal.)

In fact, several experiments show that ordinary subjects do regard at least some cases of double prevention as less paradigmatically causal than some cases involving energy/momentum transfer (e.g., Walsh and Sloman 2011). However, in what follows I focus on experiments by Lombrozo (2010) showing that people make distinctions among relations of double prevention, judging some as more paradigmatically causal than others. Lombrozo finds that subjects are much more willing to regard relations of double prevention as causal when these involve a biological adaptation, the designed function of an artifact, or an intentional relationship. Following a suggestion made in Woodward (2006), Lombrozo proposes a reason for this: like other sorts of relations of counterfactual dependence, relations of double prevention can vary in stability. Suppose that *E* counterfactually depends on *C*, where the dependence is nonbacktracking and intervention supporting. Stability has to do with the extent to which this relation of counterfactual dependence would continue to hold as other factors in the background, in addition to *C* and *E*, change. Casual observation suggests that, other things being equal, dependence relations that are more stable are, as an empirical matter, judged as more paradigmatically causal than less stable relations. This effect is present in pairs of examples not involving double prevention, but it also provides a natural explanation of Lombrozo's results.[6] In particular, as argued both by Lombrozo and in Woodward (2006), it is plausible that cases of double prevention involving biological adaptations, designed functions, and intentional actions tend to be more stable than cases of double prevention lacking these features. For example, double prevention relations involving biological adaptations tend to be more buffered against environmental disruptions (and more stable for that reason) than double prevention relations that arise, so to speak, fortuitously. Thus, the influence of the presence of lactose in the environment of *E. coli* on the expression of the enzymes involved in lactose synthesis is commonly understood by biologists as a straightforwardly causal relationship (involving "negative control"), even though this is a double prevention relation, while in Hall's bombing example, the overall dependence relation between *c* and *e* could be disrupted by any one of several non-far-fetched environmental changes, leading us to regard it as less stable and hence less paradigmatically causal. Lombrozo also suggests, more speculatively, that counterfactual relations involving energy/momentum transfer tend, ceteris paribus, to be more sta-

6. Lewis (1986) compares the unstable (or, as he calls it, "sensitive") causal relation between writing a letter of recommendation and the existence of certain people in the distant future with the much more stable relation between shooting at close range and death.

ble than many double prevention relations and that this explains why the former are judged more paradigmatically causal.

In addition to providing a unified explanation of why the presence of biological adaptations, designed functions, and intentional action (and perhaps transference) all tend to boost causal ratings, stability has an obvious normative significance, as argued both in Woodward (2006) and by Lombrozo (2010). More stable relationships tend to be more exportable or generalizable to a range of different contexts and, in a world of changing (or unknown) environmental contingencies, such relationships tend to be more useful and reliable for purposes of control. Thus, at least within an interventionist framework, it "makes sense" and is normatively appropriate for subjects to distinguish in the way that they do among relationships of counterfactual dependence in terms of their degree of stability. In other words, the interventionist framework provides a rationale for the distinctions in judgment observed in Lombrozo's experiments.

Much of the literature on causation proceeds by invoking conflicting "intuitions" about causation: some have the intuition that relations of double prevention are causal, and others do not; some think that causation requires transference or a connecting process, and others do not. Rather than engaging in an inconclusive debate about whose intuitions are "correct," an alternative approach, followed in the work discussed in this article, is to empirically investigate the factors that influence people's intuitive judgments and to ask whether these judgments and the factors that influence them have some recognizable normative rationale. Proceeding in this way allows us to link the empirical and the normative and may provide some insight into why we operate with notions of causation that incorporate certain features and not others.

## REFERENCES

Bonawitz, E., D. Ferranti, R. Saxe, A. Gopnik, A. Meltzoff, J. Woodward, and L. Schulz. 2010. "Just Do It? Investigating the Gap between Prediction and Action in Toddlers' Causal Inferences." *Cognition* 115:104–17.

Cheng, P. 1997. "From Covariation to Causation: A Causal Power Theory." *Psychological Review* 104:367–405.

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

Gopnik, A., and L. Schulz. 2007. *Causal Learning: Psychology, Philosophy and Computation*. New York: Oxford University Press.

Griffiths, T., and J. Tenenbaum. 2009. "Theory Based Causal Induction." *Psychological Review* 116:661–716.

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

Kushnir, T., and G. Gopnik. 2005. "Young Children Infer Causal Strength from Probabilities and Interventions." *Psychological Science* 16:678–83.

Lewis, D. 1973. "Causation." *Journal of Philosophy* 70:556–67.

———. 1986. "Postscripts to 'Causation.'" In *Philosophical Papers*, vol. 2. Oxford: Oxford University Press.

Lombrozo, T. 2010. "Causal-Explanatory Pluralism: How Intentions, Functions, and Mechanisms Influence Causal Ascriptions." *Cognitive Psychology* 61:303–32.
Sloman, S., and D. Lagnado. 2005. "Do We 'Do'?" *Cognitive Science* 29:5–39.
Spirtes, P., C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.
Steyvers, M., J. Tenenbaum, E. Wagenmakers, and B. Blum. 2003. "Inferring Causal Networks from Observations and Interventions." *Cognitive Science* 27:453–89.
Tomasello, M., and J. Call. 1997. *Primate Cognition*. New York: Oxford University Press.
Waisman, A., A. Cook, A. Gopnik, and L. Jacobs. 2009. "Causal Inference in the Domestic Dog (*Canis familiaris*) and Preschool Age Children." Paper presented at the 15th International Conference on Comparative Cognition, Melbourne, FL.
Walsh, C., and S. Sloman. 2011. "The Meaning of Cause and Prevent: The Role of Causal Mechanism." *Mind and Language* 26:21–52.
Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.
———. 2006. "Sensitive and Insensitive Causation." *Philosophical Review* 115:1–50.
———. 2007. "Interventionist Theories of Causation in Psychological Perspective." In Gopnik and Schulz 2007, 19–36.