Connecting Counterfactual and Physical Causation

Winston Chang (winston-chang@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Rd., Evanston, IL 60208 USA

Abstract

The study of causal judgment is dominated by two lines of research. According to one of these, what makes something a cause is that the effect's occurrence depends on the cause. Some of these theories hold that a cause increases the probability of the effect; others hold that a cause is necessary for the effect. In the second line of research, what makes something a cause is that it has a physical connection to the effect. Some of these theories hold that a cause must transmit force or energy to the effect. These two lines of research make commitments to different mental representations of causal relationships. In the present studies, participants were asked to make causal judgments about stories where the two kinds of theories make conflicting predictions. The results can be explained by neither kind of theory alone, although a hybrid model may be able to explain the results.

Keywords: causation; counterfactuals; physical causation

In our daily lives, we observe the world and effortlessly judge that one thing causes another. The process by which we make these causal judgements, or attributions, is a disputed matter among cognitive scientists. The disagreements here are not about mere surface details of the process of casual attribution. They are fundamental, and the candidate theories make very different claims about how we represent and think about causal relations.

Theories of causal attribution generally fall into one of two categories. On one hand there are theories that claim that the key to our causal attributions is that the outcome depends somehow on that event. So, for example, one such theory says that if you push a door and the door opens, what makes your push a cause of the door opening is that the latter would not have occurred without the former.

On the other hand, there are those theories that claim that our causal attributions involve an analysis of the physical processes of a system—for example, that there is a physical force transmitted from the cause to the effect. In such theories, what makes your push a cause of the door's opening is that there is a force transferred from you to the door.

In the simple case described here, both types of theories render the same prediction about the causal attribution we will make: the push caused the door to open. However, there is not agreement in all cases.

Dependency Theories

In dependency theories, what makes some event A a cause of some event B is that B's happening is contingent on A, regardless of the details of the physical processes relating the two.

In some theories, these dependencies are statistical in nature, as in Cheng's Power PC (1997) theory and Pearl's causal Bayes nets (2000). In these theories, what makes some A a cause of B is that A's occurrence results in an increased probability of B. Others theories hold that the relevant relation is a *counterfactual* dependence (Lewis, 1973). In these theories, what it means when we say "A caused B" is that A happened and B happened, and if A had not happened, B would not have happened.

Those causal theories which are based in statistical dependencies tend to be geared toward explaining *type* causal claims, as exemplified in the statement, "Drinking alcohol causes car accidents." In contrast, counterfactual causation is geared toward *token* causal claims, as in "Drinking alcohol caused this accident." There are theories which use statistical dependence to account for token causation (Halpern & Pearl, 2005; Lewis, 2004), but this paper will focus on counterfactual dependence, since it is more naturally suited to token causation. However, much of the discussion about counterfactual dependence will apply to statistical dependence as well.

Although counterfactual theories capture many of our intuitions about causal relations, they are not successful in all cases. From the philosophical literature, there is a well-known class of problem cases. It is a variety of causal overdetermination known as preemption. Consider the following story:

> *Suzy and Billy, expert rock-throwers, are engaged in a competition to see who can shatter a target bottle first. They both pick up rocks and throw them at the bottle, but Suzy throws hers a split second before Billy. Consequently Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's had not occurred, so the shattering is overdetermined.* (Hall, 2004)

In this story, there is a strong intuition that Suzy's throw caused the bottle to shatter and Billy's throw did not. But even if Suzy had not thrown the rock, the bottle still would have shattered, and so according to the counterfactual account, it is not a cause. Furthermore, the two throws have the same counterfactual relation to the bottle shattering. This would imply that they have the same causal relationship to the bottle shattering, but such a conclusion seems to miss something about how we actually make causal attributions.

Physicalist Theories

Physicalist theories of causal attribution take causal relations to be rooted in physical processes. In the philosophical literature, Salmon (1997) and Dowe (2000) have proposed that causation involves the exchange of conserved physical quantities such as energy or momentum.

Wolff (2007) has proposed the *dynamics model*, in which causal relations involve the representation of patterns of forces and a position vector. The dynamics model has a number of differences from the philosophical physicalist theories. First, it is designed to account for the meaning of not just *cause*, but also *prevent*, *enable*, and a number of other related words. Second, it is a psychological account of causal attribution, whereas the philosophical physicalist theories are not meant to be psychological—for example, Dowe (2000) seeks "to establish what causation in fact *is* in the actual world."

In physicalist theories, causation has to do with the transfer of forces or other physical quantities. So, for example, what makes my push a cause of the door opening has to do with the force transferred from my body to the door. Contrast this with the counterfactual account, in which the push is also judged a cause, but for entirely different reasons. Although knowing the physical processes might be useful for a person to deduce *that* there is a counterfactual dependence, those processes do not *constitute* a causal relation in counterfactual theory; all that matters is that the two events happened, and that if the first had not happened, the second also would not have happened.

Physicalist theories can make the correct predictions of people's causal judgments in cases of preemption, as in the rock-throwing story. In that case, Suzy's rock contacts and transfers energy or force to the bottle, whereas Billy's does not; and so Suzy's throw is considered the cause, and not Billy's.

Physicalist theories agree with counterfactual theories in most cases (as in the door opening) and can better handle cases of preemption. Although they have an advantage in some cases, they fall short in others. Consider the following story:

> *Two airplanes circle an airport, waiting to land. An air traffic controller sees on his screen that the planes are on a collision course, so he presses his radio's transmit button and instructs the pilots to change course. However, just before he does so, a saboteur cuts the power to the radio. Without power to the radio, the message is not transmitted, and the planes crash into each other.*

In my informal polling, there was near unanimous agreement that the saboteur caused the plane collision. But notice that the saboteur and air traffic controller have no physical connection to planes. Had the saboteur not intervened, there would have been a signal sent from the controller to the planes, but in fact, the two groups of objects (saboteur and controller, and the two planes) are physically isolated from each other. Cases like this, where something is considered a cause when there is a counterfactual dependence but no physical connection, are not uncommon in our everyday lives (Schaffer, 2000).

Two kinds of cause?

A theory of how we attribute causes should correctly predict our behavior in causal attribution. The two kinds of theories we have seen so far each succeed in some cases, but not all. Neither theory seems to be able to fully account for our actual causal attributions.

A third possibility is that "cause" is polysemous. It may have, for example, a counterfactual meaning and a physicalistic meaning. In most cases, it is not necessary to distinguish between them because they agree in most cases whether or not something is a cause, but when the two meanings of "cause" differ about what counts, as in the rock-throwing and airplane stories above, they compete, and one may win out over the other. This is similar to a proposal by Hall (2004) in which there are two concepts of causation.

Walsh and Sloman (2005) have investigated the possibility that both counterfactuals and mechanism can play a role in causal attribution. They conducted a series of studies on how counterfactuals and physical connections influence judgments of *cause* and *prevent*. In their studies, they found that for causes, physical connections mattered more, but for preventers, counterfactual dependence mattered more. These results, however, do not explain those cases where we call something a cause when there is no physical connection, but there is a counterfactual dependence.

The purpose of the present experiments is to investigate the possibility that there are two competing meanings of causation. One possibility is a scenario is evaluated with respect to each meaning of cause; if the two evaluation methods give different answers, each result is weighted and a weighted average is taken. If there is a simple competition between the two meanings, then each of the meanings in question should play some role in causal attributions.

Experiment 1

In the first experiment, participants were given questionnaires containing four stories, each with an event A and event B, where A was a candidate cause for B. For each story, there were four variants, which differed along two independent variables. The first variable was whether or not B counterfactually depended on A (these conditions are denoted +CFD and –CFD). The second variable was whether or not there was a physical connection between A and B (+PHY and –PHY).

Each page of the questionnaire contained a story, an illustration depicting the events in the story (shown in Figure 1), and a question asking whether event A was a cause of event B. For example, here is a story in the +PHY, +CFD condition:

> *Jim is playing with a model train set. There is a fragile house of cards on the track. There is a train that needs a push to get started, but once started, it moves on its own power. Jim pushes the train, and it moves forward. It hits the house of cards, and the cards fall down. If Jim had not pushed the train, the cards would not have fallen.*

The question for this story:

> *Was Jim's action (pushing the train) a cause of the house of cards falling down?*

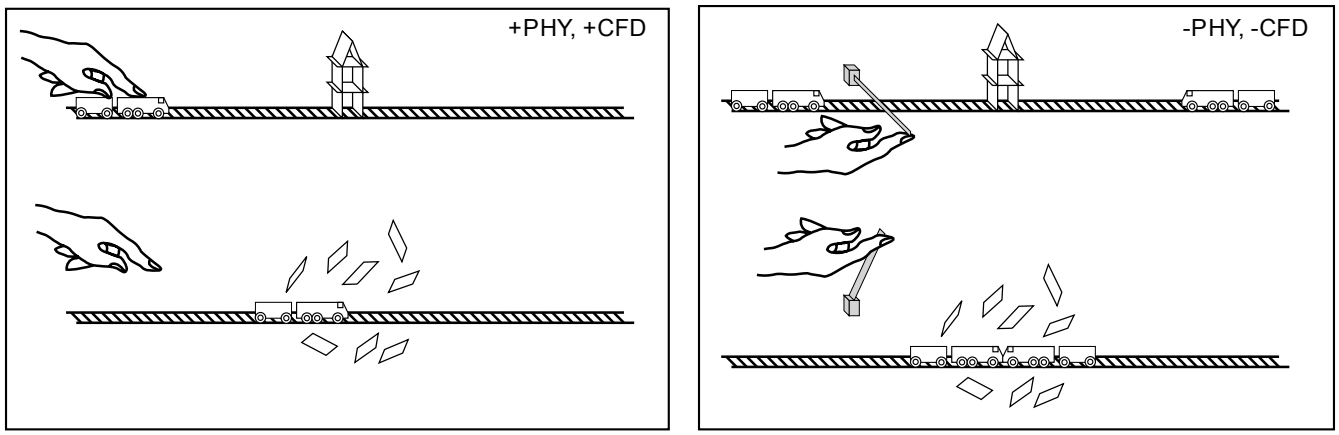

Figure 1: Examples of illustrations used in Experiment 1. On the left is a scenario where the outcome (the cards falling) has both a physical connection to and counterfactual dependence on the potential cause (Jim pushing the train). On the right, there is no counterfactual dependence because of the second train, and no physical connection that transmits force between the action (Jim lifting the gate) and the outcome. The other two conditions are a mix of these two; in the +PHY,–CFD condition, Jim pushes the train and there is a second train, and in the –PHY,+CFD condition, Jim lifts the gate and there is only one train.

Participants answered on a scale from 1 to 7, where 1 represented *definitely no* and 7 represented *definitely yes*.

The version of the story above was in the +PHY condition, so there was a transfer of force from A (Jim pushing the train) to B (the cards falling). In the –PHY condition, there is not a transfer of force from Jim to the train; the train is already coming, but there is a gate in the way that would block its path. Jim lifts the gate before the train reaches it, and the train passes through.

The version of the story above was in the +CFD condition, so there was a counterfactual dependence of B (the cards falling) on A (Jim pushing the train). In the –CFD condition, there is no counterfactual dependence of B on A; there is a second train coming from the other direction, and the two trains strike the cards at precisely the same time. The –CFD versions of this story stated that even if Jim had done nothing, the cards still would have fallen.

Forty-eight students introductory psychology undegraduates from Northwestern University participated in the experiment. They were given a packet with four stories, one in each of the four conditions.

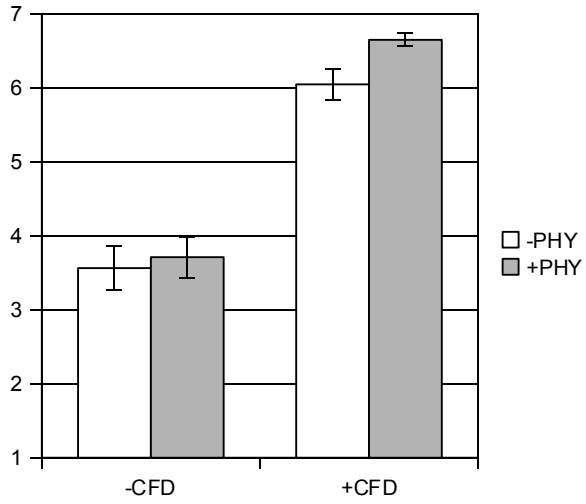

Figure 2: Causal ratings in Experiment 1.

Results

There was a significant main effect for counterfactual dependence, such that the presence of counterfactual dependence resulted in a higher causal rating, $F(1,47) = 126.89$, $p < .001$. A significant main effect for physical connection was not found, $F(1,47) = 3.06$, $p = .087$. No significant interaction between the two variables was found, $F(1,47) = 1.26$, $p = .241$.

At a first glance, the results here seem to indicate that, in making causal judgments, counterfactual dependence is far more important than a physical connection. This runs counter to the initial hypothesis, that both counterfactual dependence and physical connection have role to play. Recall that in the rock-throwing story with Billy and Suzy, the physical connection seems to override the counterfactual dependence. The structure of the rock story differed from the stories in this experiment: all the –CFD stories in this experiment were cases of symmetric overdetermination, in which two objects struck the target simultaneously; the rock story, on the other hand, is a case of preemption, where one object struck the target before the other one had arrived, and so the second object did not touch the target.

The results here also seem to run against those of Walsh and Sloman. They found that in attributing causation, the presence of a physical connection mattered more than coun-

terfactual dependence. Similar to the rock-throwing story, they variety of overdetermination used in their stories was preemption, as opposed to the simultaneous overdetermination used here.

Experiment 2

The purpose of this second experiment was to investigate how order affects causal judgments. In the previous experiment, the two objects arrived at the target simultaneously, but in this experiment, the order at which they arrived was varied.

Three of the four stories were adapted from the previous experiment.[1] There were two independent variables: physical connection, as in Experiment 1, and object order. For order, there were three possibilities: the object of interest (e.g., Jim's train) arrived at the target first (denoted 1ST), the two objects arrived simultaneously (SIM), or the other object arrived first (2ND). In this experiment, *physical connection* was a between-subjects variable and *order* was a within-subjects variable. Since this experiment was meant to investigate the effect of order in cases of causal overdetermination, counterfactual dependence was not a variable; there were two objects in all stories, and no versions of the stories had a counterfactual dependence of B on A.

Forty-eight students participated in the experiment, all of whom were recruited from the Northwestern University psychology participant pool, as in Experiment 1. Half received packets with +PHY stories, and the other half received packets with –PHY stories. Each packet contained three stories, one in each of the order conditions, and a questionnaire on the last page that asked them to explain why they gave the answers that they did. The presentation was similar to that in Experiment 1, with an illustration, story, and question on each page.

Results

There was a significant main effect for order, $F(2,92) = 105.13$, $p < .001$. A main effect for physical connection bordered on significance, $F(1,46) = 3.76$, $p = .059$. No significant interaction between the two variables was found, $F(2,92) = .51$, $p = .601$.

The results from Experiment 1 showed a much larger effect for counterfactual dependence than for physical connection. If we suppose that it were true that counterfactual dependence mattered much more than physical connection, we should make the following predictions. First, order should not play a role in the causal rating, because it does not change the final outcome. Second, since none of the six versions of each story had a counterfactual dependence, they should all elicit a low rating for the causal question. The actual results, however, contradict both of these predictions: there is a large

[1]The fourth story from Experiment 1 could not be adapted because the +PHY versions of it used a hydraulic system in which the two potential causes continuously imparted force to the target; the +PHY version of the train story, in contrast, involved an impulse from Jim, but then it moved without any transfer of force from or to him.

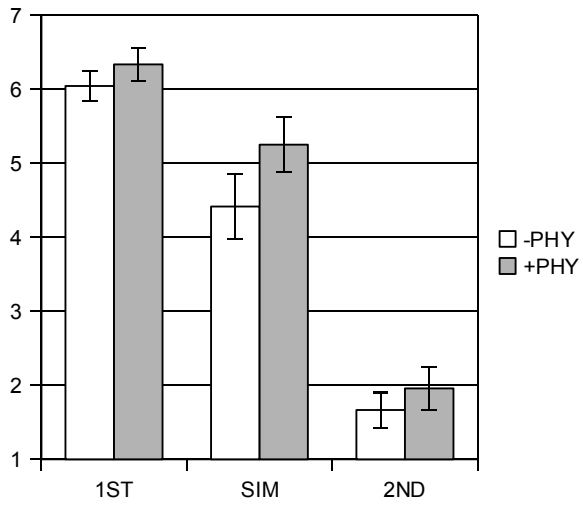

Figure 3: Causal ratings in Experiment 2.

effect for order, indicating that it is not the case that counterfactual information is the primary driver of causal attributions.

Note that there is a peculiarity in one of the conditions. In the +PHY, 2ND condition, Jim does not actually have a physical connection to the cards. There would be a connection, except that the cards have already fallen by the time Jim's train arrives. It is worth keeping this in mind when interpreting the data, although it does not appear to impact the main findings here.

One condition in particular stands out as unexplained by counterfactual theories, physicalist theories, and theories which take these systems to be in simple competition. This is the –PHY, 1ST condition. Stories in this condition received a high causal rating, which cannot be explained by a counterfactual theory, since there was no counterfactual dependence, nor by a physicalist theory, since there was no physical connection. Nor can the results be explained by a theory that has these systems competing against each other, since any competition between a "no" and a "no" should result in "no".

How is this result best explained? One possibility is that instead of having a simple competition between counterfactual and physical meanings for "cause," the scenario is decomposed into a chain of events, and each event is evaluated counterfactually or physically for who or what is responsible for the outcome. So for example, Jim is judged as responsible for the train passing through the gate based on counterfactual criteria: if he had not lifted the gate, the train would not have passed through. And his train is judged as responsible for the cards falling, because it is the one which struck and transferred force to the cards. In short, the first part is judged counterfactually and the second part is judged physically. As for why the first part should be counterfactual and the second physical, this is a question open for further study.

Conclusion

The results in these experiments are not explained by a purely counterfactual theory, a purely physicalist theory, nor by a theory that has two independent meanings for "cause" (or equivalently, two methods of making causal attributions). One way of explaining the results uses a hybrid method of making causal attribution. Instead of having counterfactual and physical analyses of the entire scenario as a whole, the story is broken down to a series of events, and each part is evaluated counterfactually and physically. If this is correct, it remains unanswered why a counterfactual analysis should prevail in some cases, and a physical analysis in others, and this issue deserves further investigation.

The task used in these studies was linguistic, and the question used was always, of the form, "Was A *a cause* of B?" It seems likely that the answers would have been different if the question was "Was A *the cause* of B?" or "Did A *cause* B?" Using these questions would help us to understand the finer distinctions of causal language; however, it is unclear whether these questions would reveal a great deal more about the mental processes underlying causal thinking.

If the suggestions here are correct, a complete theory of causal attribution may require a synthesis of counterfactual and physicalist models. Future research will help us understand how these different processes fit together.

Acknowledgements

I thank Lance Rips, Rumen Iliev, Joshua Knobe, for their helpful discussions on this research. I also thank Sirisha Yadlapati and Blayne Smith for assistance collecting data.

References

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Dowe, P. (2000). *Physical causation*. Cambridge University Press.

Hall, N. (2004). Two concepts of causation. In *Causation and counterfactuals.* Cambridge, MA: MIT Press.

Halpern, J. Y., & Pearl, J. (2005, December). Causes and explanations: A structural-model approach. part I: Causes. *Br J Philos Sci*, *56*(4), 843–887.

Lewis, D. (1973). *Counterfactuals*. Blackwell Publishers.

Lewis, D. (2004). Causation as influence. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 75–106). MIT Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Salmon, W. C. (1997). Causality and explanation: A reply to two critiques. *Philosophy of Science*, *64*(3), 461.

Schaffer, J. (2000, June). Causation by disconnection. *Philosophy of Science*, *67*(2), 285–300.

Walsh, C. R., & Sloman, S. A. (2005). The meaning of cause and prevent: the role of causal mechanism. In *Proceedings of the 27th annual conference of the cognitive science society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.