Severe Outcomes and their Influence on Judgments of Causation

Clare R. Walsh (clare.walsh@plymouth.ac.uk)
School of Psychology, University of Plymouth
Plymouth, PL4 8AA, U.K.

Ulrike Hahn (hahnu@Cardiff.ac.uk) and Lisa DeGregerio
School of Psychology, Cardiff University
Cardiff, CF10 3AT, U.K.

Abstract

The aim of this paper is to examine whether the nature of an outcome influences attributions of causation. We contrast two theories of how people make causal judgments. Counterfactual theories assume that *c* causes *e* if a change to *c* would have brought about a change to *e*. In contrast, generative theories propose that causation occurs if there is a causal process linking *c* and *e*. Both theories share the assumption that judgments of whether an event causes an outcome should be independent of the nature of that outcome. We describe an experiment showing that people give higher ratings of causation to a severe than a neutral outcome when there is no causal process linking the action and the outcome. Individuals seek causal explanations for a severe outcome more than a neutral one and when an analysis of the mechanisms fails to provide one, they are more likely to rely on a counterfactual analysis to deliver one.

Keywords: Causation; Counterfactuals; Mechanisms

Causal Attribution

Causal attributions are central to our ability to explain the world around us. Unwanted events, for example, failing an exam or involvement in an accident are particularly likely to elicit causal explanations (McEleney & Byrne, 2006). Our aim is to examine whether these unwanted outcomes also change the way we attribute causation. We will review two theoretical views of how people attribute causation. Both views assume that the severity of the outcome is independent of whether or not a particular event caused it. We then examine experimentally whether people's judgments do, in fact, reflect this independence.

Counterfactual Theory of Causation

A predominant view is that causal attribution is judged with reference to a counterfactual conditional. In other words "event *c* caused event *e*" provided that "if *c* hadn't occurred then *e* wouldn't have occurred" in the closest possible world to our own in which *c* occurred (Lewis, 1973; see Bennett, 2003 for a review). For example, I may judge that touching the hot pan caused my hand to be burnt provided that if I hadn't touched the pan, my hand wouldn't be burnt. Counterfactual theories therefore assume that causation can be inferred without reference to a causal mechanism. These theories are not without their problems. One important difficulty occurs when an outcome is over-determined, that is, when there is more than one event that is sufficient to produce an outcome and therefore undoing one will not change the outcome. However, there are variations of counterfactual theory which overcome this problem. Lewis (2000) proposes that *c* causes *e* if an alteration to *c* would have led to an alteration of *e*. Alternatively it is possible that *c* causes *e* provided that if *c* had not occurred and other causes were absent, then *e* would have occurred (Halpern & Pearl, 2001). There is evidence that people use counterfactuals in making causal attributions. For example, when the factual events in a scenario are held constant, an event tends to be judged to have a greater causal role if there is an available alternative that would have led to a different outcome (Wells & Gavanski, 1989).

Generative Theories of Causation

In contrast to counterfactual theories of causation, generative theories define causation with reference to how an outcome is brought about (Dowe, 2000; Salmon, 1984). According to these theories, causation occurs when there is a transmission of a conserved physical quantity such as energy or momentum along a causal pathway (Bullock, 1985; Salmon, 1984; Shultz, 1982) and an interaction involving an exchange of those conserved quantities (Dowe, 2000; Salmon, 1997). For example, a rock may be judged to have caused a window to break because the rock interacted with the window and transmitted energy to it. There is also psychological evidence to support the view that people think about the mechanism when making causal attributions. For example people tend to use the word 'cause' to describe a situation where something is forced towards an outcome against its initial trajectory (Wolff, 2007) and they are more likely to search for evidence about possible mechanisms than for evidence about which events co-vary with the outcome (Ahn, Kalish, Medin & Gelman, 1995). In addition, people are more likely to attribute causation to an action that is linked to the outcome by a continuous mechanism than to one that is not, even if a change to either action would bring about a change to the effect (Walsh & Sloman, 2005, 2009). Even young children are more likely to make an attribution of causation when there is a continuous mechanism linking a cause to the outcome (e.g., a rolling ball hits a jack in the box which pops) than when there is not (e.g., the ball doesn't reach the jack in the box; Bullock, 1985) and they are more likely to use information about the mechanism than spatial or temporal contiguity (Schultz, 1982). Despite extensive discussion, however, the

basis for causal attribution is still hotly debated. We next turn to a parallel debate on the role of counterfactuals and mechanisms in judgments of responsibility for doing and allowing harm to occur.

Omission Bias

There is considerable evidence to suggest that people are sensitive to the difference between actions and failures to act. People judge harm caused by actions to be morally worse than the same harm caused by a failure to act (Kagan, 1989). The difference is also reflected in many legal systems. For example, most jurisdictions allow passive but not active euthanasia (Baron, 1998). Moreover, people regret bad outcomes caused by actions more than those caused by failures to act (Byrne & McEleney, 2000; Kahneman & Tversky, 1982). And, people prefer inaction over action when both could lead to a negative outcome even if the risk of that outcome is lower following action (Ritov & Baron, 1990). Importantly, these differences also occur in judgments of causation. Actions tend to be assigned a greater causal role than failures to act when both result in the same harmful outcome (Spranca, Minsk & Baron, 1991).

The two key theories which attempt to explain this effect map closely onto our theories of causation. One explanation is that the difference between actions and failures to act depends on counterfactuals. People may be more likely to spontaneously generate a counterfactual following an action than a failure to act (Byrne & McEleney, 2000; Kahneman & Tversky, 1982). In addition, these counterfactuals can lead to different inferences. Following an action, it is possible to imagine that the actor's absence would have changed the outcome. In contrast, following a failure to act, it is possible to imagine that the actor's absence would have lead to the same outcome (Kagan, 1989) and that the actor therefore did not cause the outcome. An alternative explanation draws on generative theories of causation and relates to the fact that actions tend to linked by a continuous mechanism to an outcome whereas failures to act do not.

The two theories make different predictions for cases involving an action which interrupts a mechanism and which is therefore not linked by a continuous mechanism to an outcome. For example, if Jack jumps up and catches a ball before it hits a window then he has acted but his action is not linked by any mechanism to the window. Yet if he hadn't acted, the window would not have broken. Generative theories predict that individuals should be less inclined to attribute causation in these cases compared to when there is a continuous mechanism linking the action to the outcome. In contrast, counterfactual theories predict that people will attribute causation equally often following both types of action. Both should elicit a counterfactual that will change the outcome. Table 1 shows the predictions of mechanism and counterfactual theories for each type of scenario. Evidence suggests that people's judgments follow the predictions of generative theories, at least when outcomes are neutral (Walsh & Sloman, 2009).

Table 1: Predictions of Counterfactual and Generative theories for Attributions of Causation following Actions and Failures to Act and depending on whether there is a mechanism linking the Action to the Outcome

	Generative Theory	Counterfactual Theory
Action-Mechanism	High	High
Action-No Mechanism	Low	High
Inaction-No Mechanism	Low	Low

Outcome Severity

Theories of causation assume that judgments of whether or not an event is causal should be independent of the nature of the outcome. Counterfactual theories define causation in terms of whether a change to the outcome brings about a change to the effect regardless of what that change is. Generative theories predict that an event will be judged to be causal when there is a continuous causal process leading to an interaction with the outcome but again this judgment should not be influenced by the nature of that outcome. The aim of our study is to examine whether people's causal judgments do in fact respect this independence.

Although it is not known whether outcome severity influences causal judgments, it has been shown to influence attributions of intentionality (Knobe, 2003). In addition, it has been shown to influence a range of judgments related to causation such as responsibility and blame. Although the research examining this influence has produced very mixed results, a meta-analytic review suggests that there is a small but significant effect (Robbennolt, 2000). The primary explanation for why outcome severity should influence whether or not an individual is judged to be responsible has been termed *defensive attribution* (e.g., Shaver, 1985). According to this account, as an outcome becomes increasingly bad, individuals become more reluctant to believe that the outcome could occur to them and they become increasingly eager to attribute responsibility for the event (Walster, 1966). An alternative possibility is that the nature of the outcome may shift individuals' reliance on mechanisms or counterfactuals in making a judgment. Hence, in our experiment we used different scenarios that allowed us to test whether judgments are based on an analysis of the mechanisms present or on the counterfactual alternatives when the outcome is neutral and when it is severe.

Experiment

We generated six versions of a scenario about a boulder rolling down a cliff. We elicited judgments of whether an

actor caused a neutral outcome (the boulder falling off the cliff) or a severe outcome (the boulder knocking a man off the cliff). The scenarios also involved one of three different structures. The *action-mechanism* scenario had a mechanism connecting the action (the boulder is pushed) to the outcome. In the *inaction-no mechanism* scenario there was a failure to act (a gate on the pathway of the boulder is not closed) which changed the outcome. The *action-no mechanism* scenario involved an action (a gate on the path of the boulder is opened) but there was no mechanism connecting this action to the outcome. If causal judgments are based on counterfactuals, then individuals will give the same causal ratings to the *action-mechanism* and *action-no mechanism* scenarios. In contrast, if causal judgments are based on the mechanisms present then individuals will give the same causal ratings to the *action-no mechanism* and *inaction-no mechanism* scenarios. Hence, our design allows us to test whether people judge causation by analyzing the mechanisms present or the counterfactuals supported when the outcome is neutral and when it is severe.

Method

Participants Participants were 120 undergraduate students of Psychology aged between 18 and 24 who were recruited from the experimental participant panel at Cardiff University.

Materials, Design and Procedure There were three independent variables, action type, mechanism type and outcome. We constructed six different versions of a scenario about a boulder rolling down the edge of a cliff. Half of the versions had a severe outcome:

There is a man trapped on the cliff edge, the boulder hits him and he falls off and dies.

The other half had a neutral outcome:

The boulder rolls off the cliff edge.

These outcomes were preceded either by an action with a complete mechanism linking it to the outcome:

Marcus pushes a boulder which then rolls down a hill.

by an action which interrupted a mechanism and hence there was no mechanism linking the action to the outcome:

There is a boulder rolling toward the edge of a cliff. There is a closed gate between the boulder and the cliff edge. Marcus is parked in a car nearby and gets out of his car and opens the gate.

or an omission which also did not have a mechanism linking it to the outcome:

There is a boulder rolling toward the edge of a cliff. There is an open gate between the boulder and the cliff edge. Marcus is parked in his car nearby but fails to get out of his car and close the gate.

Participants were assigned at random to one of the six conditions in a between-subjects design. After reading the scenario, they were presented with one of the following questions depending on the outcome in the scenario[1]:

Did Marcus cause the boulder to roll off the edge of the cliff?

or

Did Marcus cause the man to die?

and they were asked to respond to each one by giving a rating on a 7 point scale from 1 referring to 'not at all' to 7 referring to 'very much so'.

Results

We carried out a 2 mechanism type (mechanism vs. no mechanism) x 2 action type (action vs. failure to act) x 2 outcome (severe vs. neutral) between-subjects ANOVA with causal rating as the dependent variable.[2] Table 2 presents the ratings of causation for each of the six scenarios. The results showed a main effect of mechanism type, $F(1,114) = 26.42$, MSE = 59.5, $p < .001$, reflecting higher ratings of causation when the mechanism was complete (mean = 5.5) than when it was not (mean = 3.1) and a main effect of action type, $F(1,114) = 16.79$, MSE = 37.8, $p < .001$, reflecting higher ratings of causation following an action (mean = 4.6) than a failure to act (mean = 2.4). Ratings of causation for severe (mean = 4.1) and neutral outcomes (mean = 3.7) did not differ significantly, $F(1,114) = 1.04$, MSE = 2.34, $p > .3$.

Table 2: Ratings of Causation following Severe and Neutral outcomes depending on the mechanism and action types

	Severe	Neutral
Action-Mechanism	5.2	5.8
Action-No Mechanism	4.1	3.5
Inaction-No Mechanism	2.9	1.9
All No-Mechanism	*3.5*	*2.7*

Outcome did not interact with action type, $F(1,114) = .27$, MSE = .61, $p > .6$, but it did interact with mechanism type, $F(1,114) = 3.47$, MSE = 7.81, 1-tailed $p < .04$. Planned comparisons revealed that when the mechanism was complete, there was no significant difference in causal

[1] They also answered a second question which is not discussed here.

[2] Although our study does not contain all 8 conditions, this analysis allowed us to test for the critical interaction between outcome and mechanism type.

ratings following severe (mean = 5.2) and neutral (mean = 5.8) outcomes, $t(38) = 1.59$, $p > .1$, however when there was no complete mechanism, causation ratings were higher following a severe (mean = 3.5) than a neutral outcome (mean = 2.7), $t(78) = 2.1$, $p < .04$.

Planned comparisons also revealed a significant difference in attributions of causation between the action–mechanism and the action-no mechanism conditions, both when the outcome was severe (means = 5.2 vs. 4.1), $t(38) = 2.22$, $p < .04$, and when it was neutral (means = 5.8 vs. 3.5), $t(38) = 4.58$, $p < .001$). There was also a significant difference between the action–no mechanism and inaction–no mechanism conditions when the outcome was severe (means = 4.1 vs. 2.9), $t(38) = 2.34$, $p < .03$, and when it was neutral (means = 3.5 vs. 1.9), $t(38) = 2.98$, $p < .01$.

Discussion

The results show the novel finding that causal ratings are influenced by the nature of the outcome. This influence depended on the causal mechanisms present in the scenario. When there was a complete causal mechanism, causal ratings were independent of the outcome. But when there was no complete causal mechanism causal ratings were higher following a severe outcome.

The result runs contrary to all existing theories of causal attribution. Yet there is a clear explanation. When there is a mechanism linking an action and an outcome, consideration of this mechanism provides a cause of the outcome. But in the absence of a mechanism linking an action to an outcome, the search for a mechanism may not deliver a cause. This may be satisfactory when the outcome is not significant, as it is not for example, when a boulder falls from a cliff. But when the outcome is significant and negative, as it is for example when a man falls from a cliff, people want to find a cause (McEleney & Byrne, 2006). Hence they may draw on counterfactuals to find one. Our account is consistent with earlier findings which show an effect of outcome severity on attributions of blame and responsibility and it suggests an alternative explanation for those results.

Our results suggest that both consideration of the mechanisms present in a situation and consideration of the counterfactuals that a situation supports may be necessary to give a complete account of individuals' causal attributions. Both play a role in causal judgments and their relative role may depend on the nature of the outcome. Support for generative theory is reflected in the finding that individuals attributed causation to an action more often when it was linked by a continuous mechanism to an outcome than when it was not. In addition, outcome information differentially influences judgments depending on the presence of a mechanism. However, generative theories cannot explain all of the results. In our study, we found moderate ratings of causation even in the absence of a mechanism suggesting that in some cases people make counterfactual based judgments of causation. Counterfactuals are also required to explain the difference between attributions following actions and omissions in the absence of a complete mechanism. This result may arise because following an omission, individuals judge that if the actor had been absent, the outcome would still be the same whereas following an action they judge that if the actor had been absent the outcome would be different (Kagan, 1989). Alternatively, the result may arise because individuals are more likely to spontaneously generate counterfactuals about actions than failures to act (Byrne & McEleney, 2000; Kahneman & Tversky, 1982). Either way a theory involving counterfactuals is required to explain the difference.

Our study replicates the omission bias showing higher judgments of causation following an action than a failure to act. Our result extends previous work in this area in two ways. It shows that the difference between actions and inactions occurs even when the action is not linked by a continuous mechanism to the outcome. Therefore the presence of a mechanism cannot be a full explanation for the difference between them. Second, our results show that the omission bias occurs not only when there are severe outcomes but also when the outcome is neutral.

Causal attributions underlie many of our judgments and decisions in everyday life (Sloman, 2005). They are also central to many legal judgments such as whether a person should be judged liable for an offence. An understanding of how and when people's judgments of causation are influenced by the nature of the outcome is therefore of great practical importance. We provide a step towards an understanding of this influence.

Our motivation for testing for potential context effects in causal judgment came from the observation that Western legal systems use counterfactuals in testing cases whereas past experimental studies (Walsh & Sloman, 2009) suggest that people use mechanisms to judge causation at least when the outcome is neutral. Of course, more than a single set of materials will be required to nail down the impact of different kinds of outcomes. However, even the results from the single contrast we present here challenge current theorizing about causation in two important ways. On a methodological level, our results suggest that we may know less about causation than previously assumed, because past studies of causal learning and causal reasoning have not systematically manipulated the nature of the putative effects; but we found patterns of causal attribution changing across outcome type. On a theoretical level, the results challenge the common assumption of all major theories that the attribution of causation is independent of the nature of the outcome.

Acknowledgments

We thank Steven Sloman for comments on this work.

References

Ahn, A., Kalish, C.W., Medin, D.L., & Gelman, S.A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299-352.

Baron, J. (1998). *Judgment misguided: Intuition and error in public decision making*. New York: Oxford University Press.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. New York: Oxford University Press.

Byrne, R. M. J. & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26* 1318-1331.

Bullock, M. (1985). Causal reasoning and developmental change over the preschool years. *Human Development, 28*, 169-191.

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

Halpern, J.Y. & Pearl J. (2001). Causes and Explanations: A Structural-Model Approach – Part I: Causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (pp. 194-202). San Francisco, CA: Morgan Kaufmann.

Kagan, S. (1989). *The Limits of Morality*. Oxford: Oxford University Press.

Kahneman, D. & Tversky, A. (1982). The psychology of preferences. *Scientific American, 246*, 160-173.

Knobe, J. (2003). Intentional action and side effects in ordinary language. Analysis, 63, 190–193.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (2000). Causation as influence. In J. Collins, N. Hall, and L.A. Paul (Eds.), *Causation and Counterfactuals.* Cambridge: MIT Press.

McEleney, A. & Byrne, R.M.J. (2006). Spontaneous counterfactual thoughts and causal explanations. *Thinking and Reasoning, 12*, 235-255.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: omission bias and ambiguity. *Journal of Behavioral Decision Making, 3*, 263-277.

Robbennolt, J.K. (2000). Outcome severity and Judgments of "Responsibility": A Meta-Analytic Review. *Journal of Applied Social Psychology, 30*, 2575-2609.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Salmon, W. (1997). Causality and Explanation: A Reply to Two Critiques. *Philosophy of Science, 64*, 461-477.

Shaver, K.G. (1985). The Attribution of Blame: Causality, Responsibility and Blameworthiness. New York: Springer-Verlag.

Shultz, T.R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*, 1-51.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*, 76-105.

Walsh, C. R. & Sloman, S. A. (2005). The meaning of cause and prevent: The role of causal mechanism. In B.G. Bara, L. Barsalou & M. Bucciarelli (Eds.) *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates (pp. 2331-6).

Walsh & Sloman, (2009). The meaning of cause and prevent: The role of causal mechanism. In submission.

Walster, E. (1966). Assignment of responsibility for an accident. *Journal of Personality and Social Psychology, 3, 73-79.*

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56*, 161-169.

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