Temporal Continuity and the Judgment of Actual Causation

Anonymous CogSci submission

Abstract

Psychological theories of actual causation aim to characterize which of multiple causes of an event is singled out as the primary cause. We present one such theory called the continuity account of actual causation. The continuity account treats events as changes of state in continuous time and traces a sequence of stage changes backwards through time from an event to its primary cause. The account is broadly compatible with the physical process view of causation, and we test it by asking people to identify the primary cause of events occurring in simple physical systems. An initial experiment confirms that root causes are more likely to be chosen as primary causes than are immediate causes. A second experiment demonstrates that root causes that have temporal continuity with the effect are preferred even when probability raising accounts would predict otherwise. The results of both experiments are consistent with the continuity account, and suggest that inferences about changes of state in continuous time may underpin an important class of actual causation judgments.

Keywords: causal inference; causal selection; singular causation; token causation; causal explanation

Around 5:27 pm on November 9, 1965, Martin Saltzman found himself trapped in a dark elevator roughly a quarter of a way up the Empire State building (Gelb & Rosenthal, 1965). The stopping of Saltzman's elevator was one of thousands of unexpected events that occurred that evening, and all of the people involved must have wondered about the causes of these events. Subsequent accounts of the Great Northeastern Blackout often trace these events back to the tripping of a relay on line Q29BD in Ontario that triggered a cascade of failures.

Identifying the cause of an event (e.g. the stopping of an elevator) requires a judgment of actual causation, also known as singular or token causation (Danks, 2017). Psychologists and philosophers have explored these judgments in detail and have developed formal models of actual causation, including models based on Bayesian networks (Halpern, 2016), force dynamics (Wolff & Thorstad, 2017) and mental simulation (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2020). Here we present and evaluate an account of actual causation that highlights the role of temporal continuity. On this account, the tripping of relay Q29BD is the main cause of the elevator stopping because the tripping initiated a continuous sequence of state changes that culminated in the stopping of the elevator.

Most accounts of actual causation are consistent with one of two broad views of causation: the counterfactual approach and the physical process approach (Hall, 2004). The counterfactual approach suggests that the tripping of relay Q29BD is a cause of the elevator stopping because if the relay had not tripped then the elevator would not have stopped. The physical process view suggests that the relay tripping is a cause of the elevator stopping because the two are linked by a physical process involving the transmission of force or energy. Both accounts need additional machinery in order to specify which among the many causes of an event is singled out as the main cause. Our continuity account fits most naturally with the physical process approach, and can be viewed as an attempt to bring out some implications of this approach for actual causation.

The continuity account relies on two core principles. First, the effect to be explained and its cause are both changes of state. The tripping of a relay qualifies as a candidate cause, but the steady-state setting of a relay does not. Our emphasis on state changes is consistent with the common view that causal relationships are relationships between events rather than facts or states of affairs, but Glymour et al. (2010) point out that most Bayes net accounts of actual causation ignore changes of state. State changes are embedded in continuous time, and a change that begins at a specific moment is typically the direct result of an event that occurred an instant before. Focusing on state changes therefore motivates a principle of temporal continuity that allows continuous causal sequences to be traced back in time from an effect to its main cause. Some previous work on actual causation highlights the importance of time (Stephan, Mayrhofer, & Waldmann, 2020), and our approach is highly compatible with work by Michotte and others suggesting that temporal information is often critical for identifying causes (Young & Sutherland, 2009). To our knowledge, however, previous work has not explored the implications of temporal continuity in the way that we do here.

The continuity account does not aspire to capture all of people's intuitions about actual causation. Like the physical process approach more broadly, it is most applicable to judgments about physical rather than social systems, and does not capture cases of causation by omission (Wolff, Barbey, & Hausknecht, 2010). To us it seems likely that judgments of actual causation rely on multiple principles that resist unification under a single heading (Hall, 2004; Danks, 2017). In focusing on physical systems we aim to characterize one paradigmatic class of judgments but acknowledge that additional approaches are needed to understand actual causation in other settings.

The following sections introduce the continuity account of actual causation and present two experiments designed to test core predictions of the account. Among previous theories of actual causation the most natural comparison is Spellman's probability raising account, which proposes that the actual cause of an effect is the cause that increased its probability to the greatest extent (Spellman, 1997). Our second experiment directly compares the continuity account with the probability raising account and we find that the continuity account provides the better account of our data.

The continuity account of actual causation

We introduce the continuity account using a scenario similar to the cover story used in the experiments. Figure 1a shows a network of particle detectors (white squares), including a special detector called the Gauge of Critical Moment (GCM). The detectors activate and turn black when they absorb a radioactive particle. Activation is transmitted across links in the network, and a detector activates if all of its input detectors are active.

Although we focus on processes unfolding in continuous time, it will be convenient to divide up the temporal dimension into intervals brief enough that there is at most one event per interval. The effect to be explained is an event that happens within one of these intervals: for example, the GCM's change of state from inactive to active. Given this setup, our theory can be formulated using a procedure that starts with the effect and steps backwards through time until the actual cause of the effect is identified.

The immediate cause of the GCM's activation must be an event that took place in the interval preceding the activation. In Figure 1a, the event immediately preceding the GCM's activation was the activation of I_C. Having identified this immediate cause, we then step backwards and identify the immediate cause of this event, and so on. The procedure terminates once we arrive at an event that has no immediate cause within the system of interest, and this cause is declared to be the main cause of the effect. In Figure 1a the main cause is the activation of R_C.

In most cases, the backward-tracing procedure just described will identify a single main cause of the effect. If the effect has no immediate cause within the system, then no main cause will be identified. Because there is at most one event per interval, the procedure can never identify more than one main cause.

The assumption that at most one event occurs per interval follows from the idea that there can be no coincidences in continuous time. If the temporal dimension is continuous it is exceedingly improbable that two events would occur at precisely the same time — in technical terms this kind of coincidence can be described as a "measure zero" possibility. If there are no coincidences, then slicing the time dimension sufficiently finely will ensure that there is at most one event per interval.

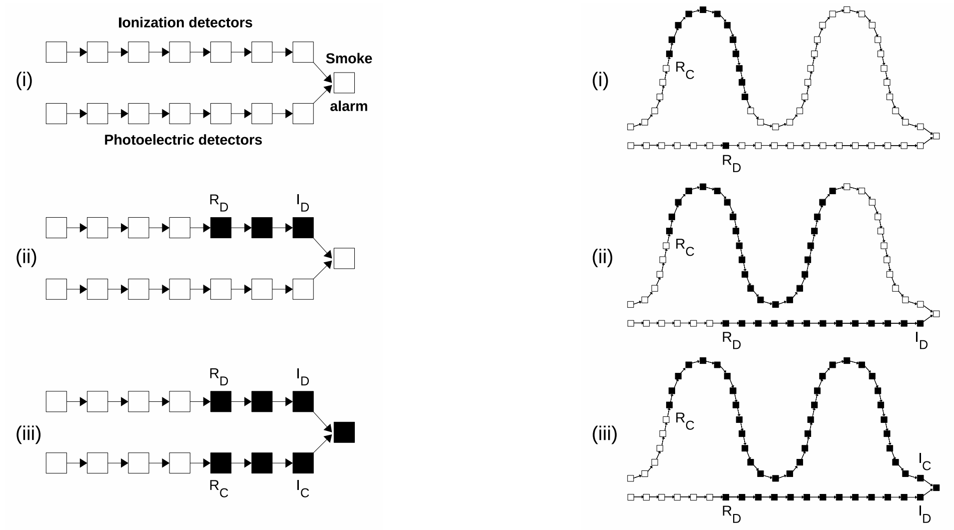
The boundaries of the causal system under consideration will depend on context. For simplicity, our discussion of Figure 1a has focused only on events internal to the particle detection network, and has attemped to characterize which of these events is best viewed as the main cause of the GCM's activation. The particle detection network, however, could also be considered part of a broader causal system that includes both the network and the network's surroundings. For example, if the network is embedded in an apparatus for carbon dating, then the spontaneous decay of a carbon atom could be identified as the immediate cause of R_C and the main cause of the GCM's activation.

Experiments

We developed two experiments to test the continuity account of actual causation just presented. Both experiments asked participants to imagine that they worked in a nuclear control room, and that their job was to monitor networks of particle detectors. Participants were told that "for each activation sequence that you see, your job is to decide what caused the activation of the GCM."

Both experiments included chains in which the GCM received input from one detector, and dual branch networks such as Figure 1a in which the GCM received input from two detectors. For chains, the continuity account predicts that participants will tend to choose the root cause of the effect (i.e. the detector that initiates the activation sequence) rather than the immediate cause (the detector that immediately precedes the GCM). For dual branch networks, the continuity account predicts that participants will tend to choose the root cause on the branch whose activation is temporally continuous with the activation of the GCM. We refer to this branch as the continuous branch, and refer to the root cause and the immediate cause on this branch as R_C and I_C respectively, where the subscript denotes "continuous." We refer to the other branch as the "delayed branch," and use R_D and I_D for the root and immediate causes on this branch. The "delay" in this naming scheme refers to the delay that occurs between I_D and the activation of the GCM. It is convenient to use the same labels for both detectors and events: for example, R_C will be used to denote both a detector on the continuous branch and the activation of this detector. This flexible use of notation, however, does not imply that actual causation involves a relationship between objects rather than events.

Our presentation of the experiments focuses on two theoretical accounts: the continuity account and the probability raising account. For completeness, though, we first discuss the counterfactual and physical process accounts of causation. Both accounts can be formulated in different ways. Here we present what we take to be the default version of each account, and argue that neither makes clear predictions about our task. The General Discussion considers ways in which these default formulations can be adjusted to better account



(a) (b)

Figure 1: (a)(i) A network of particle detectors. All components are initially inactive. (ii) Detector R_D activates and activity propagates along the top branch. (iii) Detector R_C activates and ultimately triggers the activation of the GCM, which activates only when both I_D and I_C are active. (b) An activation sequence in which (i) the top branch activates first, (ii) activation starts and finishes along the bottom branch while activation continues along the top branch, and (iii) the top branch completes its activation and ultimately triggers the GCM.

for our data.

The counterfactual account makes no clear prediction about whether root causes or immediate causes should take priority. For a chain network, if the root cause had not occurred, the effect would not have occurred, but if the immediate cause had not occurred, the effect would also not have occurred. If we consider dual branch networks and restrict attention to root causes only, then the account makes no clear prediction about whether R_C should be preferred to R_D. If R_C had not occurred, then the effect would not have occurred, and likewise for R_D.

The physical process account is similarly inconclusive. Both root and immediate causes are connected by a physical process to the effect, and there seems to be no clear reason for preferring one to the another. If we focus only on root causes, again there is no clear preference between R_C and R_D. After the activation of I_D, one can think of this detector as continuously sending activation towards the GCM which only "unlocks" the detector once I_C is also active. It follows that both R_C and R_D are connected to the effect by physical processes.

Because both of these general accounts of causation make no clear predictions, it seemed possible that people's inferences about the networks in our experiments would be highly variable and would reveal no clear trends. Our first experiment therefore explored simple cases analogous to the smoke detector example in Figure 1 with the goal of establishing whether the basic experimental procedure was viable. The second experiment focused on more elaborate dual-branch cases that aimed to distinguish between the continuity account and Spellman's probability raising account (Spellman, 1997).

Experiment 1

Experiment 1 included both causal chains and dual branch networks. For all of the dual branch stimuli, I_D occurred before R_C (Figure 2a) which means that the delayed branch completed its activation before the continuous branch began to activate. The probability raising and continuity accounts both predict that root causes should be preferred to immediate causes, and that for dual branch networks R_C should be preferred to R_D. The primary purpose of the experiment was to test both predictions.

Participants. 30 participants were recruited via Amazon Mechanical Turk and paid $3 for an 18 minute experiment.

Materials. The experiment used a customized interface built using the jsPsych library (De Leeuw, 2015). For all networks presented, participants clicked a "Run" button to observe an activation sequence. The first event in the sequence (i.e. the first change of state) always took place 5 seconds after the Run button was clicked, and the delay between successive activations along a chain of detectors was set to 100 ms. After the final event in a sequence (i.e. the activation of the GCM), all detectors became clickable after a delay of 1 second. Clicking on a detector turned its border red, and at most one detector could be selected at any time. After a sequence completed, participants could view it again if they wished by clicking a "Run again" button.

Design. The experiment included activation sequences over 15 networks (3 chains and 12 dual branch networks). Excluding the GCM, each chain and each branch of each dual network had 7 detectors. We refer to the sequences as *short*, *medium* or *long* based on the distance between the root causes and the GCM. Excluding the GCM, the short, medium and long chain sequences showed 1, 3 and 6 active detectors respectively at the end of the sequence. Within each dual branch sequence, the root causes on the two branches (R_D and R_C) were equidistant from the GCM, but these distances varied across sequences. The sequence in Figure 1a is a medium dual branch sequence (3 active detectors in each branch excluding the GCM), and short and long versions had 1 and 6 active detectors respectively per branch.

Each dual branch sequence (short, medium and long) came in four versions. The *state* version showed the delayed branch (including detectors R_D and I_D) as active from the very beginning of the sequence. The activation of this branch was therefore presented as having occurred at some indefinite time in the past, resulting in a steady state of activation. The three *event* versions all showed R_D activating 5 seconds into the sequence, and had delays of 2, 4 and 6 seconds between the activation of I_D and R_C. The predictions of the continuity account are unaffected by the delay, but we tested different delays just in case this variable affected people's responses. The continuity account also makes the same prediction about both state and event sequences, but we anticipated that the state sequences might make participants especially likely to choose R_C over R_D.

Procedure. Participants first read instructions which introduced the task and included examples of a chain and a dual branch network. They then answered three questions about the task and the detectors, and were sent back to read the instructions again if they answered incorrectly. They continued cycling through the instructions and the test questions until they answered all questions correctly.

The 15 activation sequences in the experiment proper were presented in random order. For dual branch sequences, the vertical position of the delayed branch was also randomized (Figure 1 shows the delayed branch on the top rather than the bottom). The orientation of the network (GCM on the left or the right) was randomized within participants. The prompt after each sequence was "In this sequence what caused the activation of the GCM? Respond by clicking on a detector," and participants were required to choose a single detector in response.

Results. Network orientation (left or right) and position of the delay branch (top or bottom) had no apparent effect, and we therefore collapse across these variables. Figure 2 summarizes the results for short, medium and long dual-branch sequences. The delay between I_D and R_C had no significant effect, and Figure 2 combines results for all three delays.

Although R_D, I_D, R_C and I_C all qualify as causes of the effect, Figure 2 suggests that R_C tends to be singled out as the main cause. This result can be separated into two general conclusions. First, participants were more likely to choose root causes than immediate causes. Root and immediate causes were identical for the short sequences and therefore cannot be distinguished, but the results for medium and long sequences reveal a preference for root causes. The second conclusion is that for dual branch sequences, the root cause on the continuous branch (R_C) is preferred to the root cause on the delayed branch (R_D). This result is consistent with the continuity account but also consistent with the probability raising account, because R_C guarantees the occurrence of the effect but R_D does not.

To support these two conclusions we ran two Bayesian mixed effects regression models using the *brms* package in R (Bürkner, 2017). For the first model, each response was coded as root or immediate, and all responses that could not be classified in this way (including all responses for short sequences) were discarded. We then ran a logistic regression in which stativity (i.e. state vs event) and network length (long vs medium) were included as predictors of a binary dependent variable (root vs immediate). We included a random intercept for participant to allow for statistical dependencies between multiple responses from the same participant. The posterior mean of the intercept variable was 10.54 and the 95% credible interval was $[2.30, 26.54]$, which supports the conclusion that root causes were chosen more often than immediate causes. The credible intervals for the other coefficients both included 0 (state: -2.75, $[-10.03, 1.14]$; long: 1.85, $[-1.78, 7.33]$), suggesting that the preference for root causes was not strongly affected by either stativity or network length.

The second analysis was very similar except that the binary dependent variable now indicated whether participants chose a cause on the continuous branch or the delayed branch. The posterior mean of the intercept variable was 2.42 (95% credible interval $[0.33, 4.83]$), indicating that the continuous branch was chosen more often than the delayed branch. The credible intervals for the state variable also excluded 0 (1.39, $[0.32, 2.51]$), suggesting that state sequences led to a stronger preference for the continuous branch than did event sequences. Credible intervals for both length variables included 0 (medium: 0.17, $[-0.91, 1.31]$; long: 0.11, $[-1.01, 1.21]$, where the reference level is small), suggesting that the preference for the continuous branch was not affected by network length.

It is notable that people's judgments are largely unaffected by the delay between I_D and R_C and the activation length of each sequence (short, medium or long). Both of these variables affect the time that elapses between the root causes and the effect, and our results suggest that people's judgments are not exquisitely sensitive to this sort of variation.

Experiment 2

The results of Experiment 1 are broadly consistent with both the continuity and probability raising accounts, and the goal of Experiment 2 was to distinguish between these accounts. The two make different predictions for dual branch

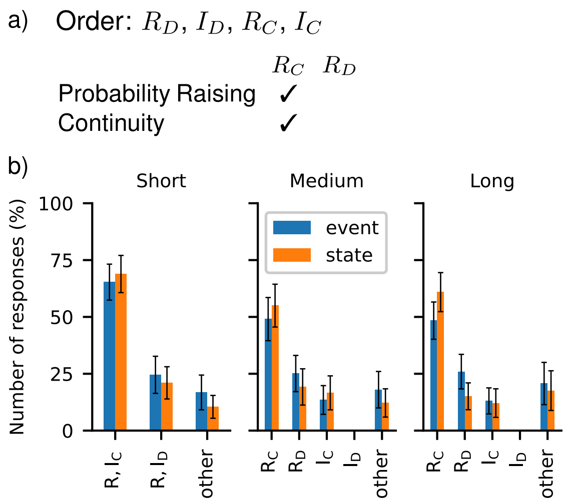


Figure 2: Experiment 1. (a) Order of the four key events for dual branch sequences. The probability raising and continuity accounts both single out R_C as the main cause. (b) Responses for dual branch sequences. The delayed branch was either inactive (event sequences) or active (state sequences) at the start of the sequence. Error bars show the standard error of the mean estimated using XXX bootstrap samples.

sequences in which the delayed branch activates after activation has already begun along the continuous branch. One such sequence is shown in Figure 1b. In cases like this, R_D guarantees the occurrence of the effect, and therefore qualifies as the primary cause according to the probability raising account. The continuity account, however, still treats R_C as the primary cause.

The materials and procedure for Experiment 2 are similar to those for Experiment 1, and we highlight only the few points of difference.

Participants. 100 participants were recruited via Amazon Mechanical Turk and paid $1 for a 6 minute experiment. Two did not submit a valid completion code, leaving 98 sets of responses for analysis.

Design. The experiment included one chain sequence and 6 dual branch sequences. Each dual branch included one straight branch and a longer curved branch, as shown in Figure 1b. The dual branch sequences included the events R_C, R_D, I_C, and I_D in three different orders shown in Figures 3a and 3b. The sequence in Figure 1b is an instance of the order in Figure 3a, because the activation on the delayed branch starts (R_D) and finishes (I_D) while activation is propagating along the continuous branch. The 6 dual branch sequences included 2 variants of each of the three orders. In one variant R_C belonged to the curved branch (as in Figure 1b) and in the other R_C belonged to the straight branch.

Procedure. The position of the curved branch (top or bottom) was randomized. The presentation order of the activation sequences and the orientation (GCM on the left or right) were randomized as for Experiment 1.

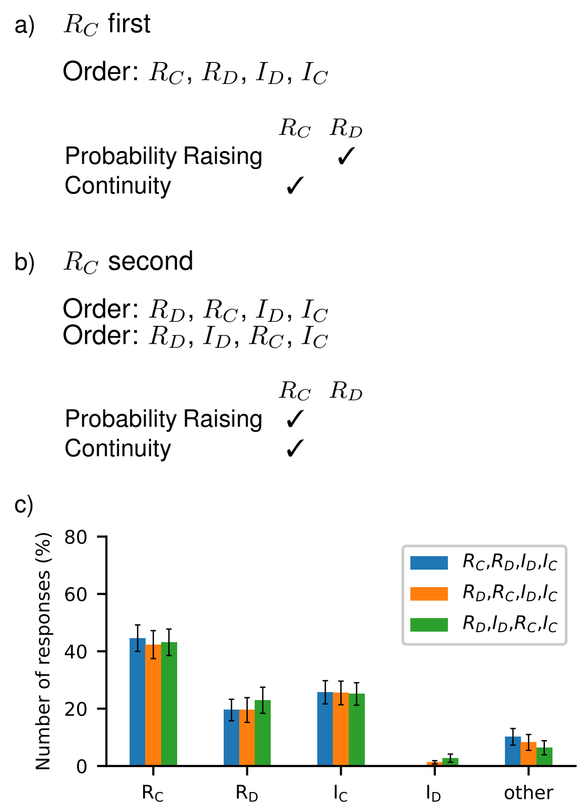


Figure 3: Experiment 2. (a) The probability raising and continuity accounts make different predictions for dual branch sequences in which R_C occurs before R_D. (b) The two accounts agree when R_C occurs after R_D. (c) Responses for the three possible orders of the key events.

Results. Network orientation (left or right) and position of the curved branch (top or bottom) had no effect and we collapsed across these variables. Figure 3c summarizes the results for the three orders listed in Figures 3a and 3b. Consistent with Experiment 1, R_C is preferred over R_D given the order $\{R_D,I_D,R_C\}$, and a similar effect was found for the $\{R_D,R_C,I_D\}$ sequences. The critical finding is that R_C is also preferred for the $\{R_C,R_D,I_D\}$ sequences, even though the probability raising account makes the opposite prediction. When referring to the orders we have dropped the fourth event because this event is always I_C.

To further analyze the data we used mixed effects models analogous to the two described for Experiment 1. The first analysis used a logistic regression in which order (i.e. the temporal order of the key events) was included as a predictor of a binary dependent variable (root or immediate). The posterior mean of the intercept variable was 5.75 and the 95% credible interval $[2.44, 10.52]$, which supports the conclusion that root causes were chosen more often than immediate causes. The credible intervals for both order variables included 0 ($\{R_D,R_C,I_D\}$: 2.11, $[-1.53, 1.21]$; $\{R_D,I_D,R_C\}$: -0.38 $[-1.78, 0.90]$, where the reference level was $\{R_C,R_D,I_D\}$), suggesting that the preference for root causes was fairly consistent across the three orders.

The second and more critical analysis used a binary dependent variable that indicated whether participants chose a cause on the continuous branch or the delayed branch. The posterior mean of the intercept variable was 2.0 (95% credible interval $[1.4, 2.65]$), indicating that the continuous branch was chosen more often than the delayed branch. The credible intervals for both order variables included 0 ($\{R_D,R_C,I_D,I_C\}$: -0.38, $[-0.95, 0.21]$; $\{R_D,I_D,R_C\}$: -0.46, $[-1.03, 0.12]$) suggesting that the preference for the continuous branch was of similar strength across all three orders. Relative to Experiment 1, the frequency with which I_C is chosen has increased in Experiment 2. This difference may reflect the increased difficulty of Experiment 2. When both branches of a dual branch structure are simultaneously active, keeping track of both root causes and the order in which they occurred is relatively challenging, which may lead some participants to fall back on the simple strategy of choosing the immediate cause that directly precedes the effect.

General discussion

We presented an account of actual causation that highlights the role of temporal continuity and described experiments that support two of its predictions. First, people tend to identify root causes rather than immediate causes as the main cause of an effect. Second, when an effect is produced by the convergence of multiple causal pathways, participants tend to identify the pathway that has temporal continuity with the effect as the primary cause.

Because the continuity account considers changes of state in continuous time, it can exploit a "no coincidences" principle to identify a single main cause of an effect. This principle applies broadly to causation in physical systems, but is less applicable to social scenarios such as voting scenarios (Livengood, 2013). Exceptions to the "no coincidences" principle are possible even for physical systems, and the continuity account can capture the illusory perceptions of actual causation that arise in some such cases (Thorstad & Wolff, 2016). For example, an 11 year old boy whacked a stick against a telephone pole at the instant at which the Great Northeastern Blackout hit his town, and ran home terrified that he had caused the power outage (Gelb & Rosenthal, 1965).

The continuity account highlights ideas such as the "no coincidences" principle that go beyond default formulations of the counterfactual and physical process accounts of causation. Key predictions of the account, however, can be reconstructed within both of these general frameworks. The counterfactual account can exploit fine-grained temporal information if each event is supplemented with a timestamp. If the effect to be explained in Figure 1 "the activation of the GCM at exactly 10 sec after the start of the trial," then the counterfactual account suggests that R_C is a cause of the effect but R_D is not, because if R_D had been different (e.g. by occurring slightly earlier or later) then the effect would still have occurred at the same instant.

As suggested earlier, standard formulations of the physical process view treat both R_C and R_D as causes of the effect in Figure 1. Our data are consistent with the possibility that people view R_C and R_D as equally valid causes, but pick R_D when forced to break the tie for reasons that may be relatively superficial. For example, perhaps R_D is preferred because the activation on the continuous branch is continuous both in space and time, and therefore more prototypical of a causal process than the activation along the delay branch. We have planned a future study that asks whether the R_C over R_D is fundamental or superficial using a task in which participants rank multiple candidate causes in order of importance, and in which ties are permitted.

The experimental paradigm we have developed can easily be extended to a more diverse set of causal systems. For example, by introducing preventive causal links, we can explore case of double prevention (e.g. cases where an active detector that would have prevented the activation of the GCM is itself inactivated by another detector). Glymour et al. (2010) point out that current work on actual causation is based largely on discussions of an "infinitesimal fraction" of the set of possible cases, but our particle detector task can support a comprehensive survey of people's judgments about actual causation.

References

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, *80*(1), 1–28.

Danks, D. (2017). Singular causation. *The Oxford Handbook of Causal Reasoning*, 201–215.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12.

Gelb, A., & Rosenthal, A. M. (1965). *The night the lights went out*. New York: New American Library.

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2020). A counterfactual simulation model of causal judgment.

Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., ... Zhang, J. (2010). Actual causation: a stone soup essay. *Synthese*, *175*(2), 169–192.

Hall, N. (2004). Two concepts of causation. *Causation and counterfactuals*, 225–276.

Halpern, J. Y. (2016). *Actual causality*. MIT Press.

Livengood, J. (2013). Actual causation and simple voting scenarios. *Nous*, *47*(2), 316–345.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323–348.

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation—–a computational model. *Cognitive Science*, *44*(7), e12871.

Thorstad, R., & Wolff, P. (2016). What causal illusions might tell us about the identification of causes. In *Proceedings of*

the 38th annual conference of the cognitive science society.
Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*(2), 191–221.
Wolff, P., & Thorstad, R. (2017). Force dynamics. *The Oxford Handbook of Causal Reasoning*, 147.
Young, M. E., & Sutherland, S. (2009). The spatiotemporal distinctiveness of direct causation. *Psychonomic Bulletin & Review*, *16*(4), 729–735.