Identifying "when" and "whether" causation: How people distinguish generation, hastening, prevention, and delay

Tianwei Gong (t-gong@ucl.ac.uk)[1,2], **Yining Hou (sherrihou@outlook.com)**[2],
Henrik Singmann (h.singmann@ucl.ac.uk)[1], **Neil R. Bramley (neil.bramley@ed.ac.uk)**[2]
[1]Department of Experimental Psychology, University College London, London, United Kingdom
[2]Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom

Abstract

Causal relationships in the real world can have diverse mechanisms with differing statistical signatures. We investigate whether people can distinguish between causes that merely change the timing of events ("when" causes) and those that bring about or prevent those events ("whether" causes). We designed experiments where the rate of an event varies over time due to one such causal influence. Events were shown in real time in Experiment 1 and as a timeline visualization in Experiment 2. Our results suggest that people are capable of identifying "when" and "whether" causes but with a distinctive pattern of confusability: People confuse Generation with Hastening; and Prevention with Delaying. We develop a Causal Abstraction from Summarizing Events (CASE) model, which explains people's judgments as mediated by their rate-change-event detection. We discuss how this line of research can be extended to study human cognition about dynamic causal influences and its relevance to real-life judgment and decision-making.
Keywords: causal learning; reasoning; time; abstraction

Introduction

Imagine you are cycling past a bus stop when, suddenly, five Bus #26s pull in one after another. You might find yourself wondering why. Perhaps the city released extra buses because of a concert or football match? Perhaps the later buses lucked out on a run of green lights, allowing them to arrive early? Or, perhaps some of these buses were delayed, causing them to pile up with later buses still running on time? Meanwhile, unless the #26 is usually an extremely frequent service, this pattern seems inconsistent with the city having cancelled some of the #26 buses today. These four possibilities represent inferences to distinct potential causal events: A cause might *generate* or *prevent* some class of events from occurring, or it might instead influence the *timing* of those events by *hastening* or *delaying* instances. Most research has focused on generation and prevention — the "whether" causes, while fewer studies have considered "when" causes, that do not affect an effect's frequency but rather its timing.

As in the traffic example above, time-sensitive causal relationships are common in reality. Sometimes, they are even the primary focus of causal inquiry. For example, medical treatments generally aim to delay decline and mortality rather than prevent it altogether; catalysts (such as enzymes) are developed to accelerate some desired process (such as the breakdown of pollutants in environmental cleanup). Even when a cause does not change the long-run frequency of its effect, altering its timing can be significant and important to us.

Are people able to use the pattern of occurrences of a variable over time and recognize the presence of "when" causes as well as "whether" causes? One possibility is that people hold a simple binary representation of causation, categorizing a cause as relatively "good" or "bad" based on its alignment with their goal. Greville and Buehner (2007) examined how people rate both "whether" and "when" causes using the same scale. Participants rated a treatment as harmful to bacteria if more bacteria died over a five-day observation compared to a non-treatment group (generating), or when the total number of deaths was the same but the treatment group experienced more deaths on Day 1 or Day 2 (hastening). Similarly, they rated a treatment as beneficial if the treatment group exhibited fewer total deaths overall (preventing) or fewer deaths early in the observation period (delaying; see also Gong & Bramley, 2024). In Greville, Buehner, and Johansen (2020), people distinguished hastening and delaying by rating the former as positive and the latter as negative, analogous to how generative and preventative causes are rated classically (Cheng, 1997). Lagnado and Speekenbrink (2010) found that when a generative cause was combined with a hastener, people's causal ratings decreased, presumably because, in their case, the hastener increased variability in the timing of the effect (Gong, Pacer, Griffiths, & Bramley, in press; Gong, Gerstenberg, Mayrhofer, & Bramley, 2023; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018). These studies demonstrate people's sensitivity to "when" causes, but they examine them under a framework similar to that used for "whether" causes. This leaves it unclear if people explicitly distinguish between the two categories.

Another possibility is that people's conceptual representation of causation is richer than can be captured with a binary framework. Given their prevalence in real world settings, people may have learned to recognize the unique patterns of rate change associated with "when" causation and developed distinct representations. This idea parallels the notion that people have intuitive theories for all sorts of physical mechanisms, and are able to reason about physical scenes by comparing observations to simulated patterns generated by these mental models (mental physical engines; Ullman, Spelke, Battaglia, & Tenenbaum, 2017; Battaglia, Hamrick, & Tenenbaum, 2013). Specifically, if a reasoner has separate representations for "whether" and "when" causation, they should form distinct impressions. As shown in Figure 1a, here we assume that the rate of effect events remains constant

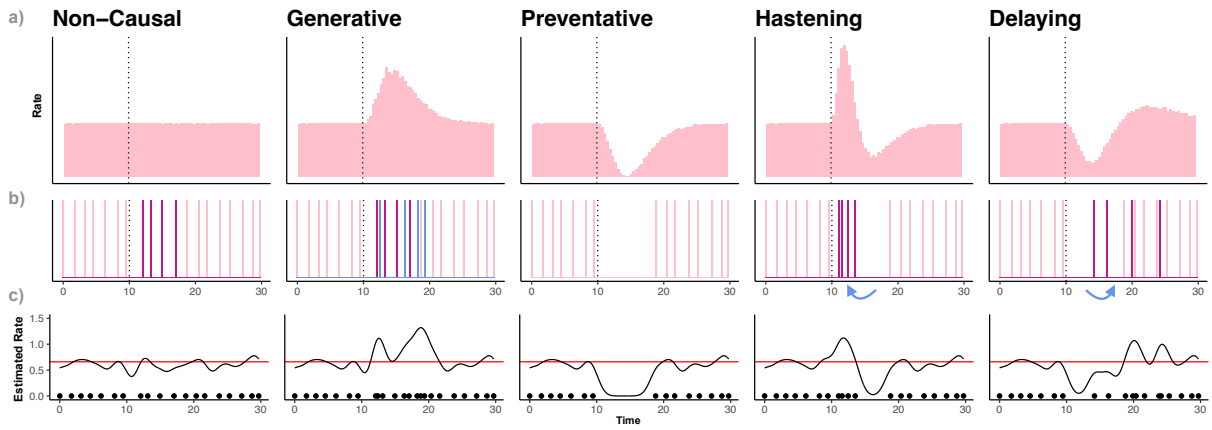

Figure 1: "Whether" and "when" causal influences. (a) Prototypes of effect-rate following different causal influences occurring at the time of the dotted line. (b) Timelines for one set of stimuli in the experiments. Effect events indicated by colored vertical lines. Purple events represent those influenced in the causal conditions. Blue events are newly generated in the Generation condition. (c) Estimated real-time rate given the sequence of events. Red horizontal line indicates baseline effect rate.

in time in the absence of external influence, while the presence of a temporary causal influence can make a difference. Specifically, one would expect a generative cause to lead to a rate increase and a preventative cause to lead to a rate decrease. In contrast, a hastening cause would initially result in a rate increase, followed by a decrease, due to a number of events being moved forward in time. Similarly, a delaying cause would initially result in a decrease, followed by an increase. This "rebound" pattern is unique and characteristic of "when" causes, as they only affect the timing rather without affecting the long-run post-cause prevalence of the effect. While the exact magnitude and timing of the rate changes can vary greatly, their qualitative patterns should be recognizable without precise expectations and could therefore serve as useful, general cues for identifying the type of causal influence one is observing.

In this paper, we examine people's ability to distinguish between four types of causes with two "whether" causes (generation, prevention) and two "when" causes (hastening, delaying). As an initial exploration into this line of research, we designed stimuli aiming for a moderate level of difficulty to avoid ceiling effects, expecting the existence of systematic errors in human data. We conduct experiments using two display formats (real-time and timeline) to deconfound perceptual and conceptual sources of errors, and two cover stories to reduce domain effects. We then develop a process-level model to explain how people make judgments. Our model combines the aforementioned summary statistics with an event detection mechanism. This model can better capture participants' judgments than a normative model, in particular accounting for the systematic errors people make.

Experiment 1: Real-time Task

Methods

Participants 120 participants (56 female, 61 male, 2 non-binary, 1 undisclosed, aged 40±12; n=60 for each cover story) were recruited via Prolific Academic and paid £2. The pre-registration (https://osf.io/6kvpz/) as well as data analysis and demos (https://osf.io/hyd7u/) are available.

Design Two cover stories were used here to limit influence of participants' prior knowledge. Participants were asked to imagine playing the role of a worker in either an airport control tower (the airplane cover story) or a strawberry greenhouse (the strawberry cover story). In both cover stories, participants were instructed that events (airplanes taking off or strawberries being picked) occurred at a relatively stable rate normally, while their task was to identify what happened when an alarm sounded. At the beginning of each trial, events occurred stably once every 1.5±0.4 s. After 10 s (i.e. at $t = 10$), the alarm lighted up and sounded, indicating that a special event might have just occurred. The trial ended after another 20 s (i.e. at $t = 30$). As shown in Figure 1b, we constructed different stimuli based on the following logic: In the non-causal condition, the interval between two effect events remained 1.5±0.4 s. For the other four causal conditions, we used a scaled gamma distribution (Gong et al., in press) with a mean of 6±3 s to control how likely events were to be influenced. This meant that the closer an expected event was to occurring at 6 seconds after the alarm, the more likely it was to be influenced. "Influence" was defined as follows: In the Hastening condition, the latency between the causal event and influenced event was halved (e.g. an event that counterfactually would have occurred 5 seconds after the causal event now occurred 2.5 seconds after); in the Delaying condition, the latency between the causal event and influenced event was doubled (e.g. an event that counterfactually would have occurred 5 seconds after the causal event now occurred 10 seconds after); in the Prevention condition, the influenced event was removed; in the Generation condition, the influenced event was duplicated appearing as an extra event occurring 6±3 s after the alarm.[1] We generated 5 unique non-causal stimuli and

[1]There are various other mechanisms we might have used to produce these patterns (see General Discussion).

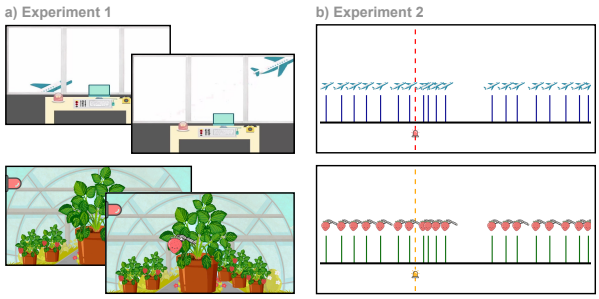

Figure 2: Cover story and stimulus example: (a) Participants in Experiment 1 watched 30s videos of a sequence of events. (b) Participants in Experiment 2 viewed the same event sequences summarized as timelines.

then created the other four types of causal stimuli by making corresponding adjustments. This resulted in a total of 25 unique stimuli in the dataset.

Procedure In each trial, participants experienced the event stream by watching a video in which the effect events were presented as quick visual animations accompanied by sound effects, occurring against a background at specific timings. After the video ended, participants answered the question, "What happened after the alarm sounded?" by selecting one of the five radio button options: No Effect, Generating, Preventing, Hastening, or Delaying. The order of 25 trials and the arrangement of radio buttons were randomized between participants.

Before starting the task, participants were instructed on the cover story including the effect events (airplanes or strawberries) and the alarm. They were not informed about any specific parameters of the causal types. Instead, they received textual instructions for each type. For example, in the airplane story, a generating cause was described as follows: "Something happened that caused some unexpected planes to take off that day. For example, perhaps an air show was taking place so extra planes were taking off to take part in that." A hastening cause was described as: "Something happened that made some planes take off earlier than they normally would. For example, perhaps a storm was forecast for later in the day and some planes that tend to be affected by weather are being rushed out of the airport earlier than they would normally have left." Participants were also told that when an event is disruptive, its impact will be temporary (lasting for a limited duration) and may affect only some airplanes/strawberries, not all, even during its effective period. There was no practice session or any feedback during the task, as the goal was to test participants' intuitions about the data patterns associated with the "when" and "whether" causes, based solely on the text instructions.

Results

Neither overall accuracy (airplane: 47%±18%, strawberry: 46%±19%; $t(118) = 0.51$, $p = .61$) nor accuracy for each stimulus type ($ps > .05$) differed between the two cover stories. Therefore, we combined the two cover stories for the subsequent analysis.

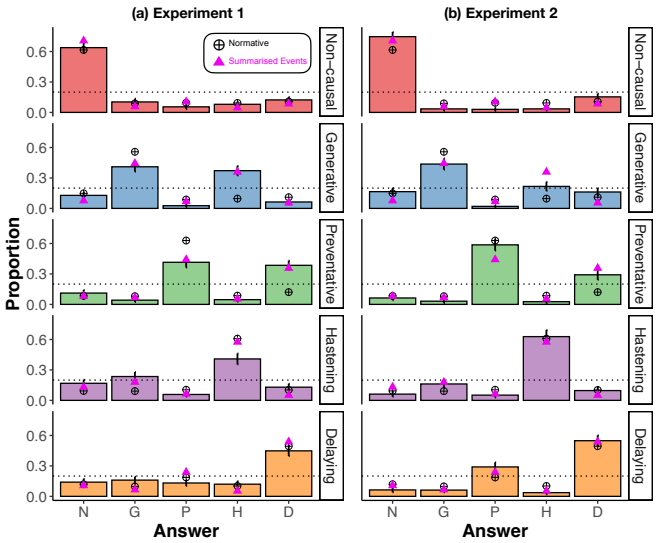

Figure 3: Participants' judgments on each conditions (bars) and results of model fitting (points).

Participants' response proportions for each option are shown in Figure 3a. Accuracy was always above chance for each type (i.e. 20%; Non-causal: 68%, $t(119) = 17.96$, $p < .001$; Generation: 41%, $t(119) = 8.62$, $p < .001$; Prevention: 42%, $t(119) = 8.15$, $p < .001$; Hastening: 41%, $t(119) = 7.65$, $p < .001$; Delaying: 45%, $t(119) = 9.62$, $p < .001$). Accuracy differed between conditions ($F(4,476) = 21.07$, $p < .001$), where Non-causal had the highest accuracy and other four conditions did not differ significantly from one another ($ps > .05$).

In the Generation condition, the proportion of choosing Hastening (37%) did not differ from the correct Generation option ($t(119) = 0.89$, $p = .38$). Similarly, in the Prevention condition, the proportion of choosing Delaying (39%) did not differ from the Prevention option ($t(119) = 0.68$, $p = .49$). It means participants systematically tended to mistake Generation as Hastening, and Prevention as Delaying.

Learning curves To further examine whether participants are capable of identifying Generation from Hastening and Prevention from Delaying, we examine whether accuracy improves during the task. For each stimulus type, we used the trial position to predict whether participants were correct or wrong on a particular trial under a mixed-effect logistic regression, with by-participant random intercepts and random slopes. Trial position was a significant predictor for Non-causal ($z = 5.15$, $p < .001$), Generation ($z = 2.39$, $p = .02$), and Prevention ($z = 3.37$, $p < .001$) stimuli, but not for Hastening ($z = 1.04$, $p = .30$) or Delaying ($z = 0.58$, $p = .56$). As Generative and Preventative stimuli in Figure 4, participants demonstrated a learning curve, becoming less likely to mistake them for Hastening or Delaying towards the end.

Discussion

Experiment 1 showed people's ability to identify different "when" and "whether" causal types. Even though they were

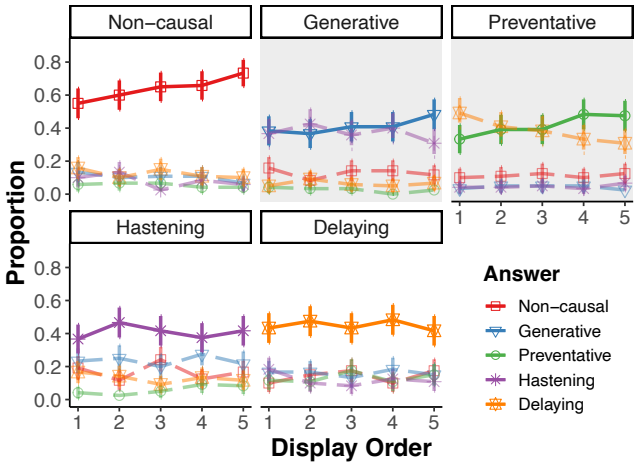

Figure 4: Changes in participants' judgments as they see a type of stimulus repeatedly in Experiment 1.

at first unable to distinguish Generation from Hastening and Prevention from Delaying, as these pairs shared similar observational patterns immediately after the alarm, their error rate decreased and the correct judgments dominated by the end of the task.

There are at least two accounts for why errors occurred in this direction rather than the reverse (i.e., mistaking Hastening as Generation or Delaying as Prevention). The first is a normative allowance for the possibility that rebound effects could occur after the end of the trial: Since delayed events could be pushed to the distant future, any observed preventative pattern might be interpreted as delaying, with the delayed events yet to appear. Similarly, observed generative patterns might be seen as hastening, with the events hastened from beyond the end of the trial. A second account is a contrast effect, which is pervasive in perception (Helson, 1964; Shapley & Reid, 1985) and has also been observed in duration estimation (Nakajima, Hasuo, Yamashita, & Haraguchi, 2014). Once participants became accustomed to the small gaps between events in the Generation condition, the later large gap may have seemed much larger, even though it merely returned to the baseline level. In other words, participants may have "hallucinated" a rebound effect even when there was no one. To further isolate the potential influence from real-time perceptual distortion, we asked participants to make judgments directly based on summarized timelines in Experiment 2.

Experiment 2: Summarized Timeline

Methods

Participants 120 participants (53 female, 67 male, aged 41±12; n=60 for each cover story) were recruited via Prolific Academic and were paid £1.50 given the shorter task duration compared to Experiment 1. Pre-registration can be found here.

Design & Procedure Experiment 2 used the same event sequences and cover stories as Experiment 1. However, for each stimulus, participants were shown a visualization of the events summarized in a timeline rather than videos (see Figure 2b). They self-paced the task by clicking to proceed to the next trial whenever they finished judging the current one. The remaining procedure and instructions were the same as in Experiment 1. There was an extra instruction page demonstrating what the timeline meant by showing a video of a non-causal event stream (not included in the formal stimulus set) paired with the summarized timeline unfolding in real time.

Results

Similar to Experiment 1, the accuracy overall (airplane: 60%±22%, strawberry: 58%±20%; $t(118) = 0.63$, $p = .53$) and on each stimulus type ($ps > .05$) did not differ between two cover stories so we combined the two for later analyses.

Participants' accuracy was higher than Experiment 1 ($t(238) = 4.92$, $p < .001$). As shown in Figure 3b, for each stimulus type, accuracy was again above chance (i.e. 20%; Non-causal: 75%, $t(119) = 23.73$, $p < .001$; Generation: 43%, $t(119) = 8.42$, $p < .001$; Prevention: 59%, $t(119) = 13.31$, $p < .001$; Hastening: 63%, $t(119) = 13.02$, $p < .001$; Delaying: 55%, $t(119) = 12.98$, $p < .001$). The proportion of participants choosing the correct option was also always higher than choosing other options ($ps < .05$). For each of the four causal conditions, the second most frequently chosen answer was the one that shared similar observational pattern right after the cause (Generation vs. Hastening; Prevention vs. Delaying).

Discussion

Experiment 2 replicated the results of Experiment 1, demonstrating that people are capable of identifying "when" and "whether" causes. This broadly supports our proposal that people have an abstract inner representation of different causal influences that is independent of the evidence format. The results confirmed the existence of real-time perception influence, as the tendency to overly mistake Generation as Hastening or Prevention as Delaying was much reduced once participants saw the full timeline rather than relying on their real-time experience of the rate changes. Nevertheless, the error patterns were systematic in both experiments, especially regarding the second most frequent judgments for each type of stimuli. This allows us to explore process models that might explain the judgment patterns across the experiments.

Modeling

Two experiments suggest that participants can use effect patterns to determine the type of causal influence. Here, we compare two models that explain how people make such judgments. Both models assume that individuals compare observed patterns with expectations associated with different types of causality. The first model proposes that these comparisons are made based on the rate experienced moment to moment. The second model suggests that the comparisons rely on higher-level cues – specifically, whether the learner experienced a "dip" or a "peak" in the rate.

Normative Model An accurate but computationally demanding approach here would be to compare *moment-by-moment* rate estimations with expectations under the different influence types. We develop a normative model that synthesizes data from the generator and updates beliefs at each moment by calculating the likelihood based on the empirical probability density function (Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018; Gong & Bramley, 2023). For example, if the synthesized data indicates that the expected effect rate at $t = 14$ (4 s after the alarm) is $[0.7 \pm 0.3, 0.9 \pm 0.5, 0.1 \pm 0.6, 0.7 \pm 0.4, 0.2 \pm 0.5]$ for N, G, P, H, D respectively, and a rate of 1.0 is estimated, the reasoner will assign the highest likelihood for this time bin to G.

To allow that participants do not know the exact causal parameters, we assume uncertainty about the timing parameters so that the events to be influenced follow distributions with a mean $\mu \sim \gamma(\mu_c = 6, \sigma_c = 4)$ after the alarm and a sd $\sigma \sim U(0, \mu)$. While the experimental stimuli model Hastening as halving the time and Delaying as doubling the time we also assume ignorance about this constraint by sampling the extent of hastening from $\beta(5,5)$ which has a mean of 0.5, and determining the extent of delaying by taking its reciprocal. These constraint relaxations go some way to make this model a more realistic normative description of the problem, and moreover help better capture human results. For this model, we fit perceptual noise with a grid search from 0.05 to 2, and fit a choice sensitivity parameter (i.e. softmax, Luce, 1959).

Causal Abstraction from Summarizing Events (CASE) Model One way to avoid computationally expensive moment-by-moment comparison is to focus on higher-level abstract influence-events ("dips" and "peaks"). Accordingly, we propose a process-level model that uses the detected influence-events as cues and compares them with *abstracted causal prototypes*. This aligns with the idea of causal abstraction (Goodman, Ullman, & Tenenbaum, 2011; Griffiths & Tenenbaum, 2009; Rehder, Davis, & Bramley, 2022), that people extract general theories from specific knowledge or observation. The prototypes for "when" and "whether" causes may have been developed via life experiences or summarized from simulated ("imagined") patterns triggered by the text characterizations of these influences. A key characteristic of these prototypes is their qualitative insensitivity to the precise influence parameters which will surely vary across situations (e.g., how long an event will be delayed if it is delayed, and what proportion of events are affected, etc). This makes the recognition of the qualitative patterns an effective and efficient path to causal influence detection in diverse and unfamiliar settings.

We define "influence-event" here as the detection of a difference from the latent counterfactual baseline expectation (which is estimable from what one observes prior to the alarm). Accordingly, the prototype of a non-causal (N) stimulus is business as usual, no influence-events are expected to be detected, since the rate of effect will hopefully continue to reflect its historical baseline. Generation (G) and Hastening (H) types are associated with an initial positive influence-event (a detectably above-baseline rate; a "peak") after the alarm, while for Hastening, this positive influence-event is followed by a negative influence-event (a detectably-lower-than-baseline rate; a "dip"). Prevention (P) and Delaying (D) types are characterized by an initial negative influence-event after the alarm, while in Delaying, this negative event is followed by a positive influence-event. As such, the learners judgment in this task can can be represented as a binary decision tree, as shown in Figure 5. Here, we use Gaussian density estimation to model how the reasoner estimates the real-time rate and apply two threshold parameters (upper and lower) to determine what influence-events are detected and consequently what kind of causal influence they are observing.

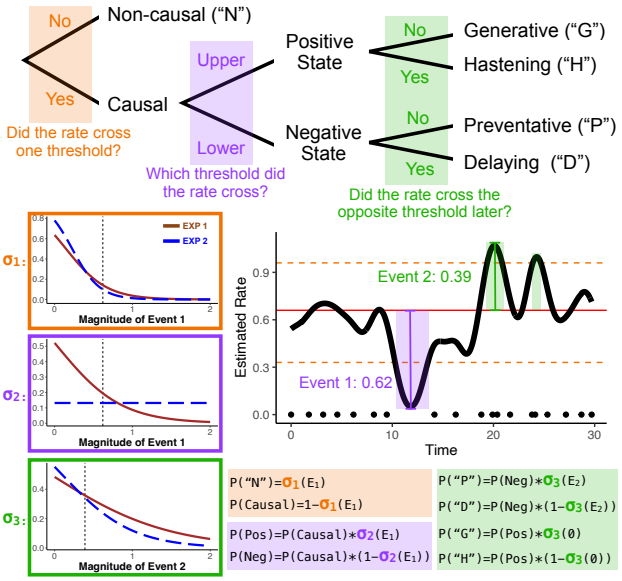

Figure 5: The CASE model. Top panel: the binary decision-making process. Bottom panel: example of how probability is calculated for a stimulus with ground truth of Delaying. Shown are the three sigmoid functions used and their fitted parameters.

The decision making process in which detected events are compared to causal prototypes is inspired by Multinomial Processing Tree (MPT) models (Riefer & Batchelder, 1988; Singmann et al., 2024). The main difference between our model and an MPT is that the parameters associated with each edge (or binary branch) in the graph are not free parameters but are determined by the magnitude of the detected events (which we define as the maximum rate difference from the baseline here; see Figure 5). More specifically, the magnitude of each event is transformed into a probability through a logistic function with two parameters: an intercept and a slope. The intercept can be understood as a general proclivity or bias towards one of the two options and the slope as the choice sensitivity for the event magnitude. This results in six parameters for three logistic functions (one for each of the three branching points).

Table 1: Model Fitting Results.

	Experiment 1			Experiment 2		
Model	BIC	CV	N	BIC	CV	N
CASE	8080	-4048	54	7116	-3602	56
Normative	8572	-4278	38	7226	-3615	52
Random	9656	-4828	28	9656	-4828	12

Modeling Results The model fitting results are shown in Table 1. We calculated the Bayesian information criterion (BIC) fit on the aggregated level and for individuals, and minimum log-likelihood on the aggregated level with a leave-one-stimulus-out cross validation (CV) procedure.[2] We include a Random baseline that has a flat 20% probability of selecting each option. For both experiments, participants judgments are better fit by the CASE model. For Experiment 1, when participants performed the task in real-time, choice sensitivities (the slopes) in CASE were weaker (see Figure 5) than in Experiment 2 (while in Experiment 2 $\sigma_2(event) = 0$ means it does not depend on the event magnitude any more). More individuals were best fit by CASE too while in Experiment 2 the gap between CASE and the normative model decreased. As shown in Figure 3, CASE also better captures the qualitative patterns in participants responses. More qualitative details are discussed in the General Discussion.

General Discussion

In this paper, we investigate how people distinguish between "when" causes (Hastening and Delaying), which only influence the timing of events, and "whether" causes (Generation and Prevention), which affect the fundamental occurrence of events. We focus on four causal types that differ in their underlying mechanisms but still share some similarities in the data patterns they produce. For example, if the rate of effects increases immediately after the cause, it could be because the cause generates extra effects or because it merely advances some future events. Two experiments demonstrated that people can distinguish these situations. They identify causal types from sequences of effects, even when the differences involve only 3–5 effect events in a timeline containing several dozen (see Figure 1b). Note that participants did not have any labeled examples, which suggests that they probably have possessed an intuitive theory about how "when" and "whether" causes differ in the patterns they produce. This finding is also consistent with the current "causal process" argument that when making responsibility judgments, people contrast the cause-removed counterfactual world with the reality and compare details on not only whether the effect would still be there, but also how the effect would look like exactly (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021).

We developed a causal-abstraction-from-summarizing-events (CASE) model to account for people's judgments. The model works by comparing summarized features of data with expectations. It further specifies the features as (positive or negative) influence-events that a learner can detect from observation. This aligns with two key empirical findings. First, people are more accurate at identifying non-causal trials compared to all other causal situations, suggesting that they may distinguish causal from non-causal cases by a primary "something happened" prediction error. Second, they tend to confuse "when" and "whether" causes when these causes share an influence-event occurring at a similar time point (e.g., both Hastening and Generation causes have a positive influence-event immediately after the cause). This could be parsimoniously explained by missing a subsequent change or hallucinating an additional one. Model fitting showed that people's judgments were better captured by the CASE model than a more normative model that attends to moment-by-moment rates. Furthermore, the CASE model demonstrated a clearer advantage in the real-time task in Experiment 1, suggesting that learners may be more likely to rely on abstract influence-events in the online setting when cognitive load is high (Lieder & Griffiths, 2020; Christiansen & Chater, 2016).

We defined the mechanisms for how delaying and hastening operate in our experiments somewhat arbitrarily. Various mechanisms likely exist in real life. For example, a delaying cause could postpone all events to a particular time point. The degree of influence might also vary based on an effect's proximity to the causal influence. For instance, events that were going to occur closer to the cause might get delayed for a longer period than events happening later. The difficulty of detecting causal influence types is also dependent on both the base rate and its regularity. All of these dimensions could be further examined in future studies.

By investigating "when" and "whether" causation, we introduce the notion of non-monotonic causation. Real-world causal influence is often dynamic; sometimes, a single cause can affect both when and how much another event occurs. For example, a policy designed to delay the outbreak of a disease may also ultimately reduce the total number of infections. Similarly, a cause may have opposite effects in the short and long run. For instance, a substance might stimulate neurotransmitter activity in the short term but reduce the total activity in the long term. Many real-life decisions involve the balance of such long-term and short-term payoffs. The current literature primarily relies on a model-free approach to studying how people handle these temporal dynamics (Caddick & Rottman, 2021; Hamou, Gershman, & Reddy, 2025; Gershman, 2025), but a causal model-based approach could also be explored, assuming that people have intuitive theories about what will generate particular non-monotonic effect patterns. By studying human recognition of dynamic causal influences, we will enrich our understanding of human representation and reasoning and gain deeper insights into how to support decision making in complex dynamic settings like healthcare.

[2] BIC's strong penalty on complex models (e.g. CASE here) may affect individual fits, for which we plan to also use CV in the future.

Acknowledgments

We thank Xintong Ji, Liza Novikova, and Yibo Wang for help with a pilot experiment of this research.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880–1910.

Caddick, Z. A., & Rottman, B. M. (2021). Motivated reasoning in an explore-exploit task. *Cognitive Science*, *45*(8), e13018.

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

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, 1–72.

Gershman, S. J. (2025). Bridging computation and representation in associative learning. *Computational Brain & Behavior*. doi: https://doi.org/10.1007/s42113-025-00242-y

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936—975.

Gong, T., & Bramley, N. R. (2023). Continuous time causal structure induction with prevention and generation. *Cognition*, *240*, 105530.

Gong, T., & Bramley, N. R. (2024). Evidence from the future. *Journal of Experimental Psychology: General*, *153*(3), 864–872.

Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N. R. (2023). Active causal structure learning in continuous time. *Cognitive Psychology*, *140*, 101542.

Gong, T., Pacer, M., Griffiths, T. L., & Bramley, N. R. (in press). Rational causal induction from events in time. *Psychological Review*.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–119.

Greville, W. J., & Buehner, M. J. (2007). The influence of temporal distributions on causal induction from tabular data. *Memory & Cognition*, *35*(3), 444–453.

Greville, W. J., Buehner, M. J., & Johansen, M. K. (2020). Causing time: Evaluating causal changes to the when rather than the whether of an outcome. *Memory & Cognition*, *48*, 200–211.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661-716.

Hamou, N., Gershman, S. J., & Reddy, G. (2025). Reconciling time and prediction error theories of associative learning. *bioRxiv*, 2025–01. doi: 10.1101/2025.01.25.634891

Helson, H. (1964). Adaptation-level theory: An experimental and systematic approach to behavior.

Lagnado, D. A., & Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal*, *3*(1), 184–195.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, 1–60.

Luce, R. D. (1959). *Individual choice behavior*. Hoboken: Wiley.

Nakajima, Y., Hasuo, E., Yamashita, M., & Haraguchi, Y. (2014). Overestimation of the second time interval replaces time-shrinking when the difference between two adjacent time intervals increases. *Frontiers in Human Neuroscience*, *8*, 281.

Rehder, B., Davis, Z. J., & Bramley, N. (2022). The paradox of time in dynamic causal systems. *Entropy*, *24*(7), 863.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339.

Shapley, R., & Reid, R. C. (1985). Contrast and assimilation in the perception of brightness. *Proceedings of the National Academy of Sciences*, *82*(17), 5983–5986.

Singmann, H., Heck, D. W., Barth, M., Erdfelder, E., Arnold, N. R., Aust, F., ... others (2024). Evaluating the robustness of parameter estimates in cognitive models: A meta-analytic review of multinomial processing tree models across the multiverse of estimation methods. *Psychological Bulletin*, *150*(8), 965–1003.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82.