Evidence from the future

Citation for published version:
Gong, T & Bramley, NR 2024, 'Evidence from the future', *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001534

Digital Object Identifier (DOI):
10.1037/xge0001534

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Experimental Psychology: General

Evidence from the future

Tianwei Gong

Department of Psychology
University of Edinburgh

Neil R. Bramley

Department of Psychology
University of Edinburgh

Author Note

We thank Marc Buehner and James Greville for providing their experimental materials, and thank Jan-Philipp Fränken, Alex Doumas, and Zachary Horne for helpful discussion. TG is supported by a Edinburgh University PPLS Scholarship. NB is part supported by a EPSRC New Investigator Grant (EP/T033967/1). Material, data, and analysis code are available at https://osf.io/h2y3g/. Correspondence concerning this article should be addressed to Tianwei Gong, Department of Psychology, University of Edinburgh, Scotland, UK. E-mail: tia.gong@ed.ac.uk.

Abstract

The outcome of any scientific experiment or intervention will naturally unfold over time. How then should individuals make causal inferences from measurements over time? Across three experiments, we had participants observe experimental and control groups over several days post-treatment in a fictional biological research setting. We identify competing perspectives in the literature: Contingency-driven accounts predict no effect of outcome timing while the contiguity principle suggests people will view a treatment as more harmful to the extent that bad treatment outcomes occur earlier rather than later. In contrast, inference to the functional form of a treatment effect can license extrapolation beyond the measurements and lead to different causal inferences. We find participants' causal strength and direction judgments in temporal settings vary with minimal manipulations of instruction framing. When it is implied that the observations are made over a pre-planned number of days, causal judgments depend strongly on contiguity. When it is implied that the observation may be ongoing, participants extrapolate current trends into the future and adapt their causal judgments accordingly. When data are revealed sequentially, participants rely on extrapolation regardless of instruction framing. Our results demonstrate human flexibility in interpreting temporal evidence for causal reasoning and emphasize human tendency to generalize from evidence in ways that are acutely sensitive to task framing.

Keywords: causality, causal learning, time, generalization

Public Significance Statement: Everyday decision making is shaped by our judgments about how the world works, such as whether a new vaccine is beneficial or harmful for our health. Since more evidence is arriving all the time, we inevitably have to make these judgments with incomplete information. Here we show for the first time that people will often use the timing of what has already happened to predict what will happen next, and then incorporate this predicted future evidence into their causal judgments. For instance, we find settings in which people take a rising trend to indicate that greater problems are to come, or a falling trend to indicate that the peak of a causal influence has already passed, leading them to make dramatically different causal judgments from the same overall case statistics. These results have implications both for understanding the sophistication of individual human causal reasoning and for understanding pinpoint the source of disagreements in public discourse around scientific evidence.

Credit Statement: Tianwei Gong served as lead for data curation, software, formal analysis, investigation, visualization, and writing –original draft. Neil Bramley served as lead for supervision, funding acquisition, writing – review and editing. Tianwei Gong and Neil Bramley contributed equally to conceptualization and methodology.

Evidence from the future

In both individual cognition and scientific practice, discovering and measuring causal effects is of central interest. Unfortunately, even with good quality experimental data and a well matched control group this can still be challenging, because genuine causal influences can take complex forms and our measurements of them are inevitably incomplete. Some effects might occur instantly and dissipate rapidly (such as from electric shocks or adrenaline injections), but others might peak later (paracetamol) grow or compound over minutes, days or years (perhaps lockdowns on covid rates, or European membership decisions on GDP). This highlights a central challenge for causal induction: To estimate the strength and direction of a novel cause, we need to decide when best to measure it. But to the extent that a treatment is truly novel, we are likely to lack the necessary mechanistic understanding to make this choice and so be forced into guesswork based on our inductive biases and whatever measurements we have.

Popular causal learning models, such as delta-P (Allan, 1980), Power PC (Buehner et al., 2003; Cheng, 1997), and Causal Support (Griffiths & Tenenbaum, 2005) contain no mention of temporal dynamics, often restricting their applicability to settings where we can assume the measurements were made at the appropriate moment to capture genuine effects. A classic scenario involves randomly assigning samples to two groups, one of which is exposed to the cause (e.g. a medical treatment) and the other of which is not. Causal judgments are assumed to be calculated based on the resulting treatment-control *contingency*, that is based on how the samples from experimental vs. control groups differ in the prevalence of the effect.

A separate line of research shows that people are sensitive to temporal information (Bechlivanidis et al., 2022; Bramley et al., 2018; Buehner & May, 2003; Buehner & McGregor, 2006; Gong & Bramley, 2023; Greville & Buehner, 2010; Greville et al., 2020; Lagnado & Sloman, 2006; Stephan et al., 2020). Event order appears to be a powerful heuristic cue to causal order which that can even override contingency information (Lagnado & Sloman, 2006). Event delays influence causal judgments (Greville & Buehner, 2010; Lagnado & Speekenbrink, 2010; Shanks et al., 1989). The temporal proximity principle, also known as *contiguity*, captures that *ceteris paribus* people make stronger causal attributions for short temporal delays than for long temporal delays (Anderson & Sheu, 1995; Grice, 1948). This applies to not only type-level judgments that reflect beliefs about which causal events cause which type of effect events (Buehner & May, 2003; Buehner & McGregor, 2006; Greville & Buehner, 2007, 2010), but also token-level judgments that reflect beliefs about which particular cause event actually caused which particular effect event (Henne et al., 2021; Ziano & Pandelaere, 2022).

Greville and Buehner (2007) built a bridge between contingency and contiguity by asking participants to evaluate the effect of treatments on bacteria survival in a day-by-day context. In contrast to contingency studies that displayed summarized outcomes, they provided participants with a sequence of numbers showing how many of the bacteria

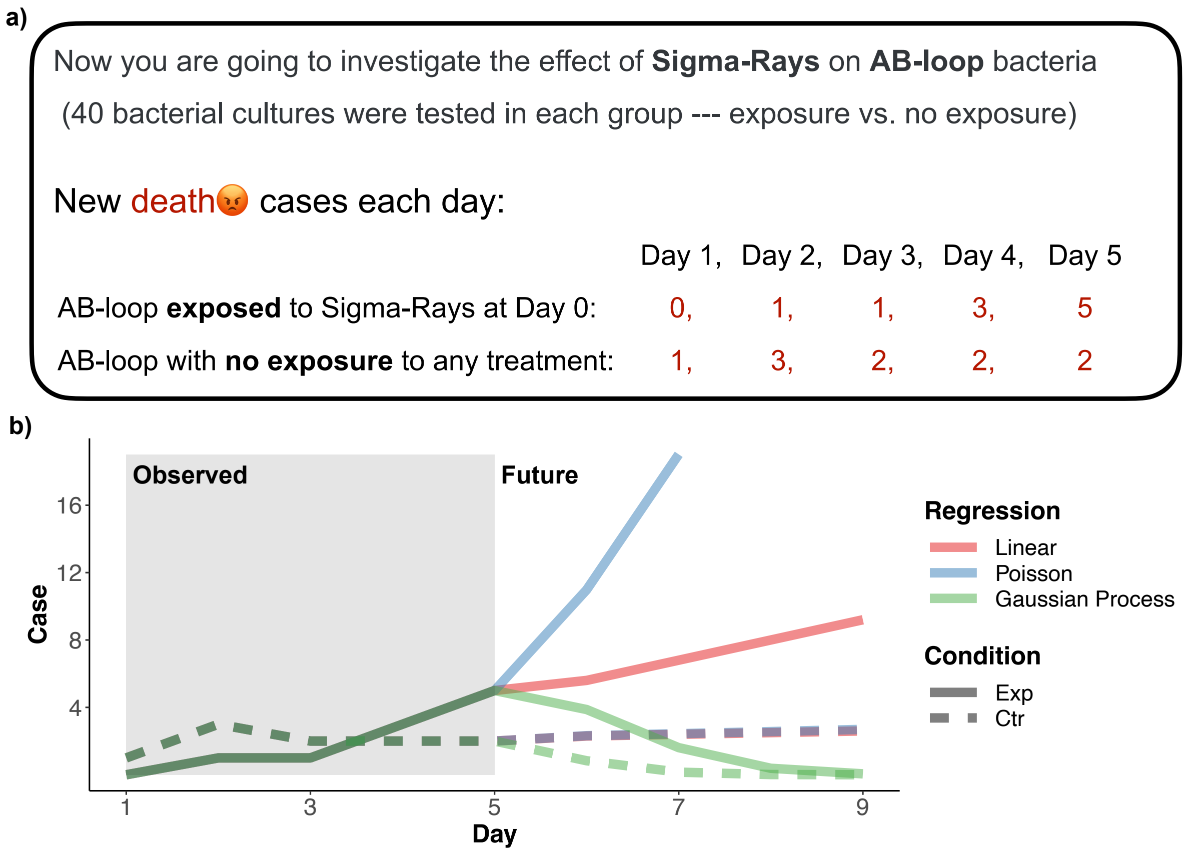

Figure 1

An example stimulus material of the current study (a) and the corresponding extrapolation results of how the new case will be in the future given different regression models (b). The Poisson regression would predict the experimental case as 0 at Day 9 due to the cumulative cases have exceeded the max sample size. The Gaussian process regression was based on RBF kernel (Schulz et al., 2017).

cultures died per day over several days. Replicating basic contingency findings, participants judged a treatment to be harmful if the experimental group had a greater total number of deaths than the control group and beneficial if the reverse was true. Meanwhile, the timing of the deaths in the experimental condition also made a difference. For the same total number of deaths, participants judged the harming effect to be greater when more of the deaths occurred on the earlier observation days and less harmful when more bacteria died on later days.

However, as aforementioned, causal dynamics could have different forms and they are unnecessary to be fully explained by the contiguity principle. In this paper, we extend on the work of Greville and Buehner (2007), showing that people not only consider temporal information, but that they can also interpret this information flexibly and adaptively. In particular, we demonstrate that instructional cues or the display format, can lead to different patterns of causal inference for the same set of observations. We demonstrate this idea with the following scenario adopted from Greville and Buehner

(2007): Imagine a biotechnology lab examines the effect of several types of radiation treatment on the survival of bacterial cultures. Bacterial cultures die naturally after a number of days, but the treatment might promote the survival of bacterial cultures (be beneficial) or kill them prematurely (be harmful). In the example shown in Figure 1a, are Sigma-Rays harmful or beneficial to the survival of AB-loop bacteria? Contingency provides no straightforward answer here since both groups have experienced the same total number of deaths by the end of the observation. According to the contiguity principle (Greville & Buehner, 2007), the treatment seems to be beneficial, potentially postponing the death of bacteria, as there are fewer deaths in the observations on days 1–3. However, one might rather suspect the treatment will ultimately prove harmful since the experimental condition has a worryingly increasing trend and most of the forty samples are still alive on Day 5. Almost any reasonable statistical model based on days 1–5 would tend to predict more death cases on days 6–9 in the experimental condition than the control condition (see Figure 1b for examples) .[1]

As demonstrated in the above example, recognizing differing *trends* across a set of measurements is another way of parsing the temporal information contained in a set of post-experimental measurements. It is possible that when making causal inferences, people consider not only the contingency and contiguity they have observed, but also whether the rates are rising or falling across the observations (and having allowed for the control condition baseline behavior) and what these suggest about the time course of the causal influence. Prediction and imagination are a key components of human cognition. Indeed, people automatically imagine possible states even if they are irrelevant to the task they have been given (Guan & Firestone, 2020). More importantly, our imagination is grounded in reality, generalizing from known circumstances to hypothetical futures and nearby counterfactual possibilities (Lucas & Kemp, 2015; McCoy & Ullman, 2019; Shtulman & Morgan, 2017). With regard to the dimension of time, people have been found to extrapolate future events by relying on the event history, even in settings set up such that each event is sampled independently (e.g. the gambler's and hot-hand fallacies Ayton & Fischer, 2004; Hahn & Warren, 2009; Szollosi et al., 2019). There is an entire research field that investigates how individuals make generalizations across contexts (Hahn & Warren, 2009; Lucas et al., 2015; Schulz et al., 2017; Zhao et al., 2022). We examine whether people further apply their generalizations from evidence to their causal judgments (Johnson et al., 2016).

To test whether people simply rely on contiguity, or also infer more complex or delayed causal influences from trends, we manipulate in three experiments what participants are told about the experimenter's stopping rule (Experiment 1), the display

[1] Of course, how many more deaths one predicts and when they will occur depends on one's specific choice of model and what inductive biases one brings to bear. In particular, the parameters of a causal generative model will depend on the the functional forms assumed for the base rate and causal effect. We do not attempt to resolve this here. The current paper mainly rely on the linear predictions, which is the common form of human generalization (Lucas et al., 2015).

Table 1
Experimental stimuli.

					Increasing		Decreasing	
	Delta-P(40)	Delta-P(15)		Total	Data	Slope	Data	Slope
A	0	0	Exp	10	0,1,1,3,5	**1.2**	3,4,2,1,0	**-0.9**
			Ctr	10	1,3,2,2,2	0.1	2,2,1,2,3	0.2
B	0	0	Exp	14	1,2,2,4,5	**1.0**	3,5,3,2,1	**-0.7**
			Ctr	14	3,3,3,3,2	-0.2	2,3,3,3,3	0.2
C	-.05	-.13	Exp	3	0,0,0,0,3	**0.6**	3,0,0,0,0	**-0.6**
			Ctr	5	2,1,1,1,0	-0.4	1,0,1,1,2	0.3
D	-.08	-.20	Exp	5	0,0,0,1,4	**0.9**	1,4,0,0,0	**-0.6**
			Ctr	8	2,2,2,1,1	-0.3	2,2,1,1,2	-0.1
E	-.10	-.27	Exp	6	1,1,0,1,3	**0.4**	2,3,1,0,0	**-0.7**
			Ctr	10	3,2,2,2,1	-0.4	1,2,2,2,3	0.4
F	.05	.13	Exp	7	0,0,2,2,3	**0.8**	3,2,2,0,0	**-0.8**
			Ctr	5	1,1,2,1,0	-0.2	1,1,0,1,2	0.2
G	.08	.20	Exp	11	0,2,2,3,4	**0.9**	2,4,3,1,1	**-0.5**
			Ctr	8	1,2,2,2,1	0	1,2,2,1,2	0.1
H	.10	.27	Exp	14	1,2,2,4,5	**1.0**	2,4,3,3,2	**-0.1**
			Ctr	10	2,2,1,3,2	0.1	1,2,2,2,3	0.4

Note: Delta-P was calculated using the sample size of 40 and 15 separately. Participants were randomly assigned to one of two stimulus lists. List 1 included the increasing version of A, C, E, G and the decreasing version of other items. List 2 included the decreasing version of A, C, E, G and the increasing version of other items.

format (Experiment 2), and the sample size (Experiment 3). We anticipate people will rely more on the trends when they focus on the possibility that more measurements are to come or that there are many more samples that have not yet been affected.

Transparency and Openness

The current work meets the Transparency and Openness Promotion guidelines suggested by the journal. We report materials and data for all experiments in Open Science Framework repository (the link is provided in the author note section). The experiments' design and analyses were not pre-registered.

Experiment 1

In Experiment 1, we investigated the impact of instructions on people's use of temporal information in causal judgments. Participants in one group were informed that there was an intended observation period that has the same length as the existing records. This was similar to Greville and Buehner (2007), thus we predicted that participants would be influenced by *contiguity* in the same manner as the previous study. In contrast, participants in another group were told that the observations would continue beyond the current records. This manipulation was intended to highlight the open future and as a result, we anticipated that participants would rely more on any *trends* in the daily case rates when making judgments.

Method

Participants. Two-hundred participants (102 female, 96 male, 1 non-binary, 1 undisclosed, aged 46 ± 13) were recruited from Prolific Academic and were randomly assigned to either the Finished (N=100) or Unfinished (N=100) conditions (see Design & Materials below). In all three experiments, participants were self-declared native English speakers located in the UK or the US and had finished at least 500 task submissions with approval rate equal or above 99%. The sample size was determined by a power analysis assuming a medium size effect of a within–between interaction and the goal of .80 power at the standard .05 alpha. Participants in all experiments received a payment of £0.50 for finishing the task. The task took around 5 minutes.

Design & Materials. We used the biotechnology cover story shown in the Introduction and manipulated three factors. Contingency (zero, beneficial, harmful) and Trend (increasing, decreasing) were manipulated within participants. As shown in Table 1, the contingency depended on the contrast of total death cases between the experimental and control groups during the observation: $P(E|C) - P(E|\neg C)$. The positive contingency is regarded as harmful and the negative contingency is regarded as beneficial. Stimuli with the same contingency could differ in their temporary trends. Increasing trends disclosed daily death cases under the experiment group with positive slopes while decreasing trends disclosed daily death cases with negative slopes. Participants were randomly assigned to one of two stimulus lists to ensure they were only exposed to either increasing or decreasing versions of the same contingency (see Table 1).

The instruction was manipulated between participants. In the *Finished* condition, participants were told that: "Bacterial cultures will be observed over a five-day period", while in the *Unfinished* group, participants were told that: "Bacterial cultures will be observed over days. The observation hasn't ended yet and the records now include Day 1 to Day 5". We predicted people would react differently to the same data given different instructions. The instructions in the Finished condition were similar to Greville and Buehner (2007), thus we predicted that participants would rely on contiguity, i.e. a decreasing daily trend with more death cases on the early days would reflect a more harmful relationship while the reverse sequence (but same overall count) suggests a less harmful relationship. In contrast, the Unfinished condition highlights the open future, and hence we predicted that participants would rely on the trend. That is, a decreasing trend should imply that in the long run there is a less harmful relationship than when there is increasing trend (which implies the cause's influence is yet to peak). Two instructions were paired with corresponding formats as shown in Figure 2.

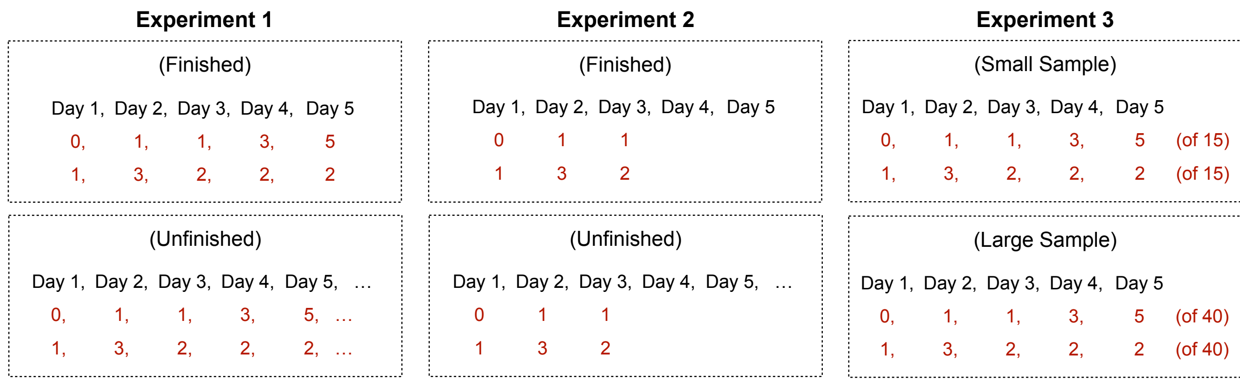

Figure 2

Stimuli displays under different conditions. Participants observed the number over days in a similar format shown in the Introduction with specific modifications illustrated in this figure. In Experiment 1 and 2, the sample size was disclosed to participants in text.

Procedure. Participants in both groups were given the biotechnology lab cover story and informed that 40 bacteria cultures were tested in each experimental or control group, the same number used in Greville and Buehner (2007). Following instruction on how to read tabular data, they were exposed to the key sentence manipulations for at least five seconds to ensure they had read them. Participants then went through 8 different pairs of treatments and bacteria. For each pair, they judge the influence of a treatment on a new kind of bacteria on a 7-point scale (-3=definitely beneficial;-2=probably beneficial; -1= perhaps beneficial; 0=not sure; 1=perhaps harmful; 2=probably harmful; 3= definitely harmful).

Results

A three-way mixed ANOVA Analysis was performed. As shown in Figure 3, there was a main effect of contingency ($F(2, 198) = 293.73$, $p < .001$, $\eta_p^2 = .60$). Pairwise comparison showed that the difference between each pair of contingency levels was significant (zero–beneficial: $t(198) = 14.72$, $p < .001$, $d = 0.69$; zero–harmful: $t(198) = 11.43$, $p < .001$, $d = 0.44$; harmful–beneficial: $t(198) = 20.79$, $p < .001$, $d = 1.13$ after Bonferroni adjustment). There was no main effect of Trend ($F(1, 198) = 0.32$, $p = .57$) or Instruction ($F(1, 198) = 0.02$, $p = .89$).

Importantly, there was a interaction between Trend and Instruction ($F(1, 198) = 14.42$, $p < .001$, $\eta_p^2 = .07$). As shown in Figure 4, decreasing trends were judged as more harmful than increasing trends in the Finished condition (simple effect: $t(198) = 3.09$, $p = .002$, $d = 0.27$), replicating the contiguity effect (Greville & Buehner, 2007). In contrast, increasing trends were judged as more harmful than decreasing trends in the Unfinished condition ($t(198) = 2.29$, $p = .02$, $d = 0.20$), indicating a trend effect. The other two-way or three-way interactions non-significant ($ps > .05$).

To check whether the interaction effect originates from the instruction manipulation

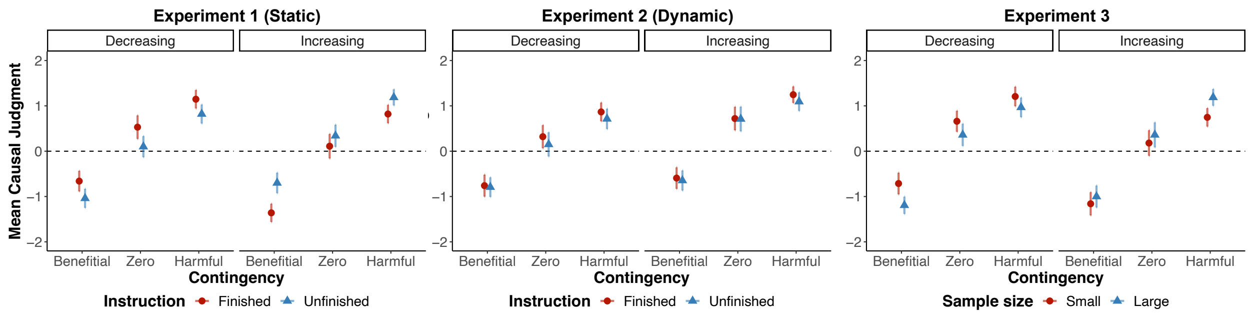

Figure 3

Means of causal judgments under different contingency and experimental conditions. Participants judged the influence of treatment on a scale from -3 (definitely beneficial) to 3 (definitely harmful). Dashed lines indicates the middle level when it is not sure whether the treatment was harmful or beneficial to the survival of the bacteria cultures. Error bars indicate 95% confidence intervals.

or the visual format difference (the dots in the Unfinished condition), we conducted a supplementary experiment (N=200) by only keeping the visual format differences between two groups (see https://osf.io/34529 for more details). Both groups were exposed to an instruction that was relatively neutral "The observation has happened for five days so far. The records now include Day 1 to Day 5". In contrast to Experiment 1, there were no any interaction effects ($ps > .05$). This suggests that the effect of the manipulation in Experiment 1 resulted from the instruction text itself.

Experiment 2

Experiment 1 showed that people not only consider contiguity when processing temporal information, but can also be sensitive to the trend, with this seemingly depending on how the choice of when the observations are made is framed. Experiment 2 investigated whether the tendency to rely on trends rather than contiguity can occur in other situations. Instead of the static display in Experiment 1, we used the dynamic display where participants click a button to reveal the data sequentially day-by-day (Soo & Rottman, 2020). This dynamic display not only reflects the reality that temporal data really are collected over time, but also reflects a setting often used in the previous research that has found people anticipate the future data based on what they have seen so far (Ayton & Fischer, 2004; Hahn & Warren, 2009; Szollosi et al., 2019). Therefore, we predicted that this real-time mode would trigger participants to anticipate the future, and so will likely rely more on the trends than contiguity when making causal judgments.

Method

Participants. Two-hundred participants (93 female, 104 male, 1 non-binary, 2 unenclosed, aged 43 ± 13) were recruited from Prolific Academic and were randomly

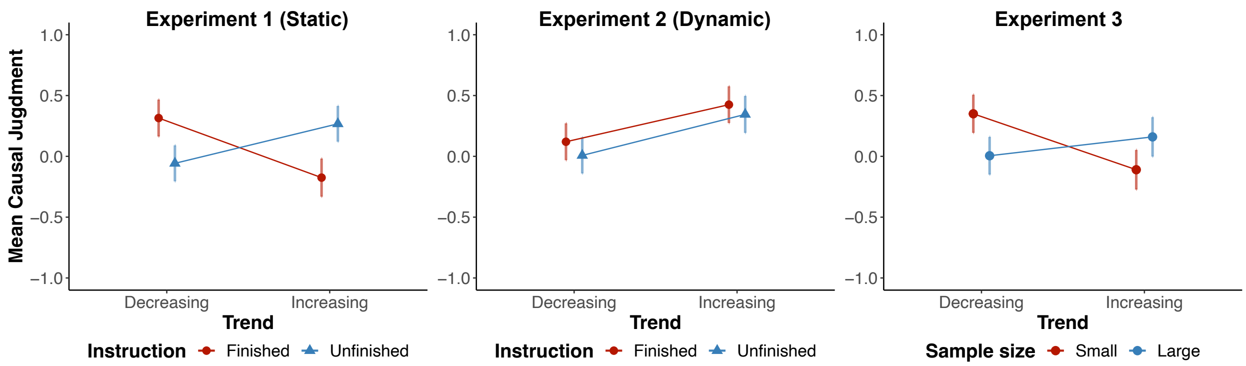

Figure 4

Means of causal judgments under Decreasing vs. Increasing trends across experimental conditions. Participants judged the influence of treatment on a scale from -3 (definitely beneficial) to 3 (definitely harmful). Error bars indicate 95% confidence intervals.

assigned to either the Finished (N=100) or Unfinished (N=100) conditions (see Design & Materials below).

Design & Materials. The experimental design and materials were similar Experiment 1. We retained the instruction manipulation but differing from Experiment 1, both groups experienced the evidence sequentially (Figure 2). Each time participants clicked on the "show the next day" button, the the next observation was revealed. Once all data had been revealed, participants in the Finished condition were prompted that "the bacterial experiment is now completed" while participants in the Unfinished condition were prompted that "the bacterial experiment continues, and you have seen the existing records".

Results

Similar to Experiment 1, there was a main effect of contingency ($F(2, 198) = 184.65$, $p < .001$, $\eta_p^2 = .48$; pairwise comparison: zero–beneficial: $t(198) = 12.61$, $p < .001$, $d = 0.63$; zero–harmful: $t(198) = 7.26$, $p < .001$, $d = 0.28$; harmful–beneficial: $t(198) = 16.48$, $p < .001$, $d = 0.91$ under Bonferroni's adjustment). There was a main effect of Trend ($F(1, 198) = 9.97$, $p = .002$, $\eta_p^2 = .05$), but no main effect of Instruction ($F(1, 198) = 0.95$, $p = .33$) or any two or three-way interaction effect ($ps > .05$). In contrast to Experiment 1, participants under both Finished or Unfinished instructions tended to rely on trend to make judgments. That is, they judged increasing trends as more harmful than decreasing trends in spite of their lower contiguity.

Experiment 3

Experiment 1 and 2 investigated the contextual factors that could influence how people utilize the temporal information. We found that participants exhibited a tendency to rely on the contiguity of deaths in the treatment condition when they experienced the data under a static display with an instruction indicating that the observation had ended. In contrast, when the uncertain future was emphasized, through either the instructions or

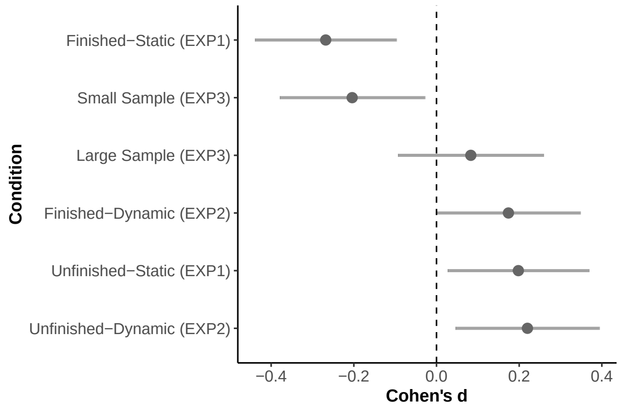

Figure 5

The Cohen's d effect size pooled out from Increasing-Decreasing simple effect tests in different conditions across experiments. Negative values mean participants prioritized contiguity over trend, while positive values mean participants prioritized trend over contiguity. Error bars indicate 95% confidence intervals of Cohen's d estimates.

by use of a dynamic display, participants tended to rely on the trend. Experiment 3 investigates a more fundamental feature of how people contextualize count data: the total sample size. A small total sample size means that participants have observed the majority of the outcomes (so there is little left to extrapolate about). A large sample leaves many cases unresolved (in our setting, many bacterial cultures that are still alive) and thus leaves more room for participants to speculate about the future.

Method

Participants. Two-hundred participants (121 female, 79 male, aged 45 ± 12) were recruited from Prolific Academic and were randomly assigned to either the small-sample (N=100) or large-sample (N=100) conditions (see Design & Materials below).

Design & Materials. The experimental design and materials were similar Experiment 1, except that instead of manipulating the instructions, we now manipulated the information of sample sizes. Participants in the *Small-sample* condition were told that both experimental and control groups tested 15 bacteria cultures, while participants in the *Large-sample* condition were informed that both groups tested 40 bacteria cultures, the same as Experiment 1 and 2 (see Figure 2). We did not include any instruction on how long the observation has lasted or whether the observation had ended at Day 5 (i.e. "Finished"or "Unfinished") in this experiment.

Results

As Experiment 1 and 2, there was a main effect of contingency ($F(2, 198) = 201.25$, $p < .001$, $\eta_p^2 = .67$; zero–beneficial: $t(198) = 15.71$, $p < .001$, $d = 0.74$; zero–harmful: $t(198) = 9.53$, $p < .001$, $d = 0.35$; harmful–beneficial: $t(198) = 20.03$, $p < .001$, $d = 1.09$ after Bonferroni's adjustment, Figure 3). There was no main effect of Trend

($F(1,198) = 0.92$, $p = .34$) or Sample ($F(1,198) = 0.09$, $p = .76$). Most importantly, as in Experiment 1, there was an interaction between Trend and Sample ($F(1,198) = 5.19$, $p = .02$, $\eta_p^2 = .03$). As shown in Figure 4, decreasing trends were judged as more harmful than increasing trends in the Small-sample condition (simple effect: $t(198) = 2.29$, $p = .02$, $d = 0.20$). The Large-sample condition showed the reverse pattern although the simple effect test was insignificant ($t(198) = 0.93$, $p = .35$, $d = 0.08$).

We can better understand the influence of temporal information in three experiments by summarizing the effect of Increasing-Decreasing simple effect tests in Figure 5. Here, a negative effect size means participants prioritized contiguity over trends, while the positive effect size means participants prioritized trends over contiguity. Participants' consideration differed across conditions. They tended to follow contiguity when the instructions indicated that the observation had ended (Experiment 1) or the data revealed the state of the majority of the samples (Experiment 3). In contrast, they showed a tendency to extrapolate the trend when they were told that the observation has not finished yet (Experiment 1) or experienced the data sequentially (Experiment 2).

General Discussion

Decades of work has studied how people learn causal relationships but it is still not clear how temporal information shapes causal inferences. Rather than exposing people to prepackaged atemporal tabular data, we here provided sequences of daily observations of an experimental and control condition. These are both more ambiguous but more informative than a simple snapshot of outcomes, since they contain information about the time profile of the causal influence (and hence whether the effect has been adequately captured by the available measurements). The mortality scenario we used here showcases this since, with a long enough time window, all the bacterial samples will naturally die meaning that there is no truly neutral time at which to compare experimental and control groups. This equifinality is a common feature of real world questions about causal effects but one that is rarely highlighted in causal cognition research.

We constructed trajectories in which new death cases after treatments increased or decreased over time. We found that participants robustly used the contingency information (Buehner et al., 2003; Cheng, 1997; Griffiths & Tenenbaum, 2005). Beyond this, they used the temporal information and used it in a malleable way. Participants judged a treatment to be more harmful if more samples died in the early days in the experimental condition, consistent with the contiguity principle found in previous studies (Buehner, 2006; Greville & Buehner, 2007; Pacer & Griffiths, 2012). However, this only happened when participants saw the data in a static format and were either told that the observation had finished (Experiment 1) or that the total sample size was so small that they had seen the most of the potential data by day 5 (Experiment 3). On the other hand, more deaths on the later days could indicate a increasing trend that would seem to herald more experimental-condition deaths in the near future. To the extent that people "play out" these possible futures in their mind, we thus expected them to draw a quite different conclusions in these situations. If people rely on the trend rather than the contiguity to make judgments, they would conversely think of high numbers of early deaths and concomitantly lower later deaths as evidence of a beneficial effect. Indeed, we found that people relied on the trend when they were informed that the observation had not ended (Experiment 1) or experienced a dynamic format where the data were revealed sequentially (Experiment 2). They showed a similar, albeit non-significant, tendency when it was emphasized that the time of death for most samples was unknown at the time of the final measurement (Experiment 3). These effects consistently occurred regardless of whether the contingency information suggested the cause to be harmful, beneficial, or non-causal. As such, this report is the first to show the boundary conditions of contiguity in case-based causal learning.

We here showed that, when utilizing temporal information, people are sensitive to the wider context (here cued by the cover story, presentation format and sample size). Whether strength judgments reflected generalization beyond the data depended on the extent that the context and the available measurements implied that all the relevant causality had been captured in the provided observations. However, it remains unclear how each factor influences the underlying cognitive process. For example, the instruction and visual format may influence different aspects. It is possible that instructions tend to influence the learner's prior expectation about causal delays, while visual formats tend to influence their use of the data: When participants are informed that the experiment ends on Day 5, they may tend to interpret this as signaling that the relevant causal influences will tend to dissipate within 5 days (else the experiment has been poorly constructed), resulting in a strong expectation for that any causal effects will be captured in the observation window. On the other hand, when participants experience the evidence in a dynamic format, they may spontaneously anticipate the future irrespective of instruction, and utilize this anticipated data to make judgments. In cases where participants are informed that the experiment continues after Day 5, they may additionally form a prior belief that the causal influence could take more than 5 days to fully manifest, and thus deliberately try to anticipate the future and summarize this with their causal judgment. Moving forward, research could employ Bayesian computational models to analyze the influence of these factors on different components of inference, i.e. in identifying the true context (tapping into priors about the relevant causal mechanisms) and interpreting the evidence (calculating appropriate likelihoods). Future work could also attempt to delineate between the more automatic component processes like involuntary extrapolation of sequences from more deliberative processing like a context-driven choice of how to interpret evidence.

One practical implication of this study is its demonstration that instructional framing influences how people interpret the data they are shown. Participants in Experiment 1 drew different causal conclusions from the same evidence depending on only a very minimal instruction manipulation. This means that providing accurate context as well as data is vital for accurate scientific communication (Soyer & Hogarth, 2012). Another key question for the future work is how people make stopping decisions when actively monitoring the outcome of their own or others' interventions or experiments. Efficient information sampling is of practical importance to cognition, since learners must balance the rewards and costs by making sensible stopping and task switching decisions (Callaway et al., 2022; Gong et al., 2023; Yu et al., 2014). This becomes even more critical in the kinds of dynamic contexts and complex causal effects that are ubiquitous in everyday life (Anvari et al., 2022; Coenen et al., 2019).

Conclusion

Across three experiments, we examined the boundary conditions of contiguity in causal inference. We found that people treated early post-intervention case levels as more important than later ones only if the majority of outcomes were subsequently observed or if they had been informed that the observations had been deliberately terminated. If told the observations would continue, or if experiencing the data sequentially, they instead focused on the trends and anticipated future evidence and concomitantly different and even reversed causal effects. Our work shows that human causal learning is not only generically sensitive to temporal information around measurements of causal effects but also to the generalizations licensed by the context in which they are measured.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

Anderson, J. R. & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition*, *23*(4), 510–524.

Anvari, F., Kievit, R. A., Lakens, D., Pennington, C. R., Przybylski, A. A., Tiokhin, L., Wiernik, B. M. & Orben, A. (2022). Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*.

Ayton, P. & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, *32*, 1369–1378.

Bechlivanidis, C., Buehner, M. J., Tecwyn, E. C., Lagnado, D. A., Hoerl, C. & McCormack, T. (2022). Human vision reconstructs time to satisfy causal constraints. *Psychological Science*, *33*(2), 224–235.

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.

Buehner, M. J. (2006). A causal power approach to learning with rates. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *28*(28).

Buehner, M. J., Cheng, P. W. & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1119.

Buehner, M. J. & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology Section A*, *56*(5), 865–890.

Buehner, M. J. & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, *12*(4), 353–378.

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Lieder, F. & Griffiths, T. L. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, 1–14.

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

Coenen, A., Nelson, J. D. & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*(5), 1548–1587.

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

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

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. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, *139*(4), 756–771.

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.

Grice, G. R. (1948). The relation of secondary reinforcement to delayed reward in visual discrimination learning. *Journal of Experimental Psychology*, *38*(1), 1–16.

Griffiths, T. L. & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Guan, C. & Firestone, C. (2020). Seeing what's possible: Disconnected visual parts are confused for their potential wholes. *Journal of Experimental Psychology: General*, *149*(3), 590–598.

Hahn, U. & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review*, *116*(2), 454–461.

Henne, P., Kulesza, A., Perez, K. & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*, 104708.

Johnson, S. G., Rajeev-Kumar, G. & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology*, *89*, 39–70.

Lagnado, D. A. & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451–460.

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

Lucas, C. G., Griffiths, T. L., Williams, J. J. & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

Lucas, C. G. & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, *122*(4), 700–734.

McCoy, J. & Ullman, T. (2019). Judgments of effort for magical violations of intuitive physics. *PloS One*, *14*(5), e0217513.

Pacer, M. & Griffiths, T. L. (2012). Elements of a rational framework for continuous-time causal induction. In N. Miyake, D. Peebles & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 833–838).

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M. & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79.

Shanks, D. R., Pearson, S. M. & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*(2), 139–159.

Shtulman, A. & Morgan, C. (2017). The explanatory structure of unexplainable events: Causal constraints on magical reasoning. *Psychonomic Bulletin & Review*, *24*(5), 1573–1585.

Soo, K. W. & Rottman, B. M. (2020). Distinguishing causation and correlation: Causal learning from time-series graphs with trends. *Cognition*, *195*, 104079.

Soyer, E. & Hogarth, R. M. (2012). The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, *28*(3), 695–711.

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

Szollosi, A., Liang, G., Konstantinidis, E., Donkin, C. & Newell, B. R. (2019). Simultaneous underweighting and overestimation of rare events: Unpacking a paradox. *Journal of Experimental Psychology: General*, *148*(12), 2207–2217.

Yu, E. C., Sprenger, A. M., Thomas, R. P. & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*(2), 268–282.

Zhao, B., Lucas, C. G. & Bramley, N. R. (2022). How do people generalize causal relations over objects? a non-parametric bayesian account. *Computational Brain & Behavior*, *5*(1), 22–44.

Ziano, I. & Pandelaere, M. (2022). Late-action effect: Heightened counterfactual potency and perceived outcome reversibility make actions closer to a definitive outcome seem more causally impactful. *Journal of Experimental Social Psychology*, *100*, 104290.