THE UNIVERSITY *of* EDINBURGH

THE UNIVERSITY *of* EDINBURGH

Causal induction in time

Tianwei Gong

Thesis submitted for the degree of
Doctor in Philosophy (PhD)

Department of Psychology
University of Edinburgh

August, 2023

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Tianwei Gong*)

To my family and friends

Abstract

Causes require time to propagate their effects. We can see stars at night because of the light they emitted hundreds of years ago. We can smell the fragrant aroma of baking bread because heat gradually changed the structure of the food, emitting particles that traveled on the breeze. In this thesis, I investigate how people use temporal information to make causal inferences. I propose a rational framework for causal induction based on continuous-time evidence, examine human performance in passive and active continuous-time causal learning tasks, and develop bounded rational accounts that can offer explanations for human causal judgments and intervention strategies.

Chapters 2 and 3 review previous theoretical frameworks on causal induction, and empirical work on the role of time in causal induction, respectively. Chapter 4 develops a rational framework for processing temporal evidence. It provides an explanation for how delays shape human causal induction and accounts for human causal judgments across seven different temporal causal learning tasks. Chapter 5 and Chapter 6 test how people passively or actively learn causal structures based on events unfolding in real time. I found people are capable temporal causal learners who successfully identify structures that involve generative and preventative relationships, as well as acyclic and cyclic connections. Nevertheless, the computational demands of normative learning could easily exceed human capacity. People's causal judgments align better with an algorithm that approximates the normative solution via a simulation and local summary statistics scheme, suggesting the reliance on structurally local computation and temporally local evidence. People's intervention decisions align better with a resource-rational model that emphasizes a balance between expected information and expected inferential complexity when choosing interventions. Chapter 7 shows that when given a limited period of observation, people not only focus on existing data, but also consider future possibilities, relying on extrapolated data to make inferences. This demonstrates the unique "continuing" feature of time, and

how generalization plays a role in the utilization of temporal information. Chapter 8 synthesizes the findings of this thesis and proposes future research directions of causal learning in temporal contexts.

Lay summary

Finding the causes of events could be very important in daily life. Imagine you are traveling to a new place and your skin gets itchy sometimes. How would you figure out the reasons behind that? Is it because of trying new food, new drinks, sleeping in a new bed, or using new shower gels? You bought pills from the store, but the locals also recommend a special herb paste. How would you test them on yourself to determine which one can efficiently remove the itchiness? In everyday life, we face this reasoning problem, where, like scientists, we need to analyze the evidence we have or conduct experiments to collect evidence. However, scientists have large, independent samples to test, such as testing a new medicine on hundreds of patients. In contrast, we have limited subjects: often the only "patient" is ourselves, and we have to test on ourselves several times at different time points to test our beliefs. This requires us to carefully think about the details of what happens in the timeline and extract useful information to make judgments. This thesis investigates how people use time information to learn the relationships between events.

By considering the orders and delays between events, people can identify the causes of target events, just like the skin itchiness example above. This includes generative causes that produce target events, as well as preventative causes that remove the target events. When asking about the relationships between two types of target events, A and B, people can identify whether A produces B, B produces A, or A and B can produce each other and have a loop relationship. This suggests that causal relationships can be learned from events unfolding in the timeline in daily life, rather than solely relying on data from scientific labs.

We can write computer algorithms to accomplish the same task of judging the relationships between events. To ensure finding the correct cause, these algorithms work like detectives, examining all the details and laying out all the possibilities that could explain what has happened. Consequently, this requires a huge amount of calculation. In contrast, people tend to ignore some details and only focus on a few possible explanations. Although the accuracy of human performance is slightly lower than that of computer algorithms, the way they calculate is much easier and faster. In other words, they sacrifice some accuracy

but save the energy. This suggests that people know how to use their limited time and limited cognitive powers to arrive at more efficient solutions.

In summary, this thesis explores the mental processes of using time information to reason about causal relationships. It contributes to our understanding of human cognition, connects to the philosophy of science and scientific education, and sheds light on issues in computer science in the quest for more human-like algorithms.

Acknowledgements

Neil Bramley, I consider myself an extremely lucky person to have worked with you on this incredibly interesting topic over the past four years. As a psychology student with zero modeling experience, you taught me how to build computational models through your patient guidance and hands-on demonstrations. You are a true master of the proximity-zone theory, for always knowing the "right" challenges, trusting in my abilities, and providing enormous supports and encouragements, even when I doubted myself. Although doing research can evoke a rollercoaster of emotions, your mentorship has transformed it into a delightful video game-like journey of ups and downs. I will cherish how much fun we've had during this journey, from navigating complex mathematical equations to hunting for funny examples for my presentations. While I now have a sense of personal growth, I am equally grateful for the knowledge, skills, and personality traits I can still learn from you. Thank you for being an outstanding supervisor.

Tobias Gerstenberg, your research introduced me to the field of causal reasoning. Thank you for crafting many papers that not only feature excellent experiments but are also incredibly enjoyable to read, for consistently publishing detailed R code for your projects, which I shamelessly borrowed, and for offering numerous valuable suggestions regarding the content in Chapter 6 of this thesis.

M Pacer, your work is a treasure trove of insights. It serves as a constant reminder of the depth and breadth that our research topic can encompass. Your PhD thesis has been a direct inspiration for Chapter 4 of this thesis and has significantly influenced my writing style. Thank you for meticulously documenting your intelligent thoughts at that time. It is truly an honor to have the opportunity to collaborate with you.

Ralf Mayrhofer, thank you for collaborating on Chapter 6 and for the insights you provided long before I joined the project. You spotted the errors in the mathematical formulas, which I know require a profound understanding of the entire project.

Dave Lagnado, Marc Buehner, Tom Griffiths, and Ben Rottman, thank you for your pioneering work on causal cognition, particularly regarding the relationship between time

and causality. It's astonishing to consider, ten or even twenty years ago, how advanced those studies were. Also, thanks so much for your kindness and valuable feedback during our interactions.

Zach Davis and Simon Stephan, thank you for your amazing studies, which are highly relevant to this thesis, and for your encouragement during (virtual) conferences; nothing is more helpful than that during those particular moments.

Anonymous reviewers, it is hard to believe how much my manuscripts improved during the reviewing process. I often think you are more familiar with my studies than I am. Thanks for all the challenges. They are well-deserved.

Chris Lucas and Rob McIntosh, thank you for teaching me so much in another research project that is not directly included in this thesis. It has significantly broadened my understanding of cognitive science.

Zach Horne and Alex Doumas, thank you for challenging me during the annual review meetings and providing me with so many suggestions. Also thanks for hosting the activities and seminars in Edinburgh that were always engaging and cozy.

Philipp Fränken, thank you for being my "unofficial" personal tutor. Thanks for your patience in listening to and analyzing my problems, even when I didn't know exactly what the problem was.

Bonan Zhao, thank you for being another personal tutor for me. Thanks for generously sharing information with me, from living in Edinburgh to job hunting, and for introducing me to such a lovely flat to live in during these years, as it significantly determines whether a PhD student can survive or not.

Stephanie Droop, Chentian Jiang, Simon Valentin, Victor Btesh, Tadeg Quillien, and Aba Szollosi, it is fascinating to have you in the same lab. Thank you for the wise and thought-provoking questions you raised in the lab meetings, for being engaged during my presentations and laughing in time when it was supposed to be a joke, for the gentle feedback on my posters and manuscripts, and for sharing your interesting thoughts about causality and learning. Thanks for being cool people. I have learned so much from you, both personally and professionally.

Fahd Yazin, thank you for asking me to explain MCMC and ABC. It was my first time to explain algorithmic logics. I am sure I didn't do a good job, but your intelligence compensated for it and made it a good experience.

School of PPLS, thank you for providing me with such a precious opportunity to pursue a funded PhD career here.

Andrew Shtulman, thank you for trusting me to be a research assistant based on a cold email. Thanks for revising my manuscripts when I had no idea how to write academic papers. Your research is always so inspiring, and I keep learning from it.

Ting Jiang, Xuefei Gao, and Jian Li, thank you for being my undergraduate supervisors. I have been thinking how different the topics I currently work on are compared to that time, but now I realize how closely they can be related.

Yuan Meng, I remember the time when I was a second-year undergraduate, we sat on the floor outside a laboratory and you talked about the blicket experiments and WebPPL. The passion poured out from your eyes. I didn't feel the same passion back then, but I certainly do now.

Xinxin Zhu, it is such an honor to have shared this journey with you as we started our PhDs and submitted our theses at the same times. Also thanks for many boxes of grapes you shared along the way.

Xiaomeng Zhang, thank you for letting me "secretly" work in your office during the pandemic, and for the time you, me, and little cute Nature spent together in my second year of the PhD, which is often referred to as the hardest year. Thanks for helping me practice my first talk, when I realized a good researcher would understand topics outside their field.

Ruomeng Zhu, your outlook on life influences me to be much less anxious. Thank you for the time we spent walking on the Meadows.

Saya Hinata, thank you for being a keen observer of the world and sharing your insights with me. You will be a great social psychologist.

Josiah King and Umberto Noe, thank you for being my bosses during my teaching assistant job. Apart from learning how to demonstrate, I learned so much about statistics through this experience. Ian Hajnosz, Otto Jutila, and Wei Li, my colleagues, I miss the wonderfully busy teaching weeks we spent together.

Huanhuan Yin and Wei Xu, thank you for the time we worked (and chilled) in the same office space. I love the occasional bets about whether we can stay until the building closes at nights, and I know you always make it. Yi Yang and Tong Xie, time is short to get to know you well, but every time we have a conversation, I feel I learn something about life.

Liyu Gao, what a miracle that we met each other at an airport abroad and realized we work in the same building at the same university, and our office desks are so close to each other. Thank you for the jogging time, which keeps me energetic.

Adri Kovacsics, your trust and kindness gave me the courage to study abroad. Thanks for that day we met at the fourth floor of BNU canteen and discovered almost no food, when we were too young to wake up early.

Sarah Kegerreis, your piano lessons are what help me get through every week. You have shown me that good pedagogy doesn't just exist in psychology textbooks. Thanks for teaching me improvisation skills, which are now helping me in many aspects of life.

Zhaonan Chen, thank you for being my go-to person for testing the pilot experiments, and always letting me know the benchmark of human performance. Your brain helps so much for my PhD. I know it is doing the same wonderful job in yours.

Yawen, Xiaomeng S, Shiyao, Qing, Mengxin, Ruijie, Minghui, Yunlan, Irina, and Xiao, thank you for the years we know each other, and for keeping in touch regardless of the physical distance around the Earth. Xinjie, Xiaoyu, Hao, and Shilin, although we haven't been in touch for many years in between, thank you for hosting me during my writing-up trip and for the time we shared our different young adulthood experiences.

Mom and Dad, as the laziest person in the family, I am always amazed by how hard-working you both are at your jobs, all while being cool, engaged in so many activities, including parenting me. Thank you for all your love and support.

Published and submitted articles

Chapter 5

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

(Materials, data, and analysis code available at https://osf.io/q8n72/)

Gong, T., & Bramley, N. R. (2020). What you didn't see: Prevention and generation in continuous time causal induction. In S. Denison, M. Mack, Y. Xu, & B. Armstrong (Eds.), *Proceedings of the 42th annual conference of the cognitive science society (pp. 2908–2914)*.

Gong, T., & Bramley, N. R. (2021). Learning preventative and generative causal structures from point events in continuous time.*Proceedings of the Causal Inference & Machine Learning workshop at 35th Neural Information Processing Systems conference*.

Chapter 6

Gong, T., Gerstenberg, T., Mayrhofer, R., & Bramley, N. R. (2023). Active causal structure learning in continuous time. *Cognitive Psychology*, *140* (4), 101542.

(Materials, data, and analysis code available at
https://github.com/tianweigong/time_and_intervention)

Chapter 7

Gong, T., & Bramley, N. R. (2023b). Evidence from the future. *PsyArXiv. Accepted at Journal of Experimental Psychology: General*.

(Materials, data, and analysis code available at https://osf.io/h2y3g/)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Time is the school in which we learn,
Time is the fire in which we burn."

Delmore Schwartz

We all learn how to navigate the journey of life. Throughout our human existence, we have embarked on a profound quest to uncover the secrets of hunting, farming, cooking, and healing. In this pursuit, we have explored both the external world and our internal selves, discovering ways to maintain physical well-being, attaining financial stability, and nurture our mental health. The challenge lies in discovering what truly "works", a challenge intricately intertwined with our understanding of the objective world's causal structures. We cultivate our understanding of these causal models through passive observations of dynamic phenomena or active interventions that yield observable changes.

Prior to the evolution of science or the advent of formal scientific education, humans possessed an inherent ability to absorb knowledge about the world. Each passing moment brings forth a myriad of occurrences, and comprehending the causal mechanisms underlying these events necessitates the assimilation of information derived from our experiences. Given that all things unfold within the framework of time, understanding how individuals acquire knowledge of causal structures from temporal events becomes not only a question of individual survival but also a testament to the collective success of humanity.

Throughout the long history of causal learning studies, participants have often been presented with evidence independently gathered from numerous entities, typically organized in tabular formats or bounded by clearly defined trials, mirroring the controlled environments of scientific laboratories. In contrast, ordinary individuals encounter events in their daily lives that occur repeatedly within a limited sphere of entities, unfolding in a continuous temporal flow without

distinct experimental trials or demarcations. Consequently, the process through which laypeople learn causal relationships within such a dynamic temporal environment remains relatively unexplored.

In this thesis, I explore the computational theories and empirical knowledge of how people learn causal structures from events unfolding in continuous time. It includes questions surrounding the *learning* of different types of causal relationships when they intertwine over time, as well as how individuals strategically *intervene* in time to enhance their understanding of causal structures. To simulate real-world situations, I set up online causal learning tasks that allows participants to observe evidence or actively intervene in real time, and subsequently making causal judgments.

To gain deeper insights into the empirical patterns observed, I build computational models that serve as valuable tools for analysis. My research question is approached under the guideline of rational analysis. I firstly build a computational-level model to describe the optimal solution of the task and then contrast this with more computationally tractable and cognitive plausible approximations. By comparing model predictions and human performance, I try to illustrate how people efficiently abstract useful evidence from rich information that happens rapidly over time, how they make trade-offs between evidence's informativeness and complexity.

The structure of this thesis is as follows:

Chapters 2 and 3 review previous empirical work, highlighting factors that influence human causal judgments, including the order of events, delay lengths, predictability, and invariance. Chapter 4 develops a rational model framework for processing temporal evidence. It provides a unified account of why short, expected, and unvaried delays shape human causal induction, but also reveals the computational challenges of normative inference. I show the framework accounts for human causal judgments across four previous datasets and three new datasets covering a variety of causal relationships (generative, preventative, acyclic, cyclic) and scenarios (multiple short observations, one extended observation).

The complexity of causal induction from temporal dynamics, demands that learners consider their cognitive limitations and find more efficient and approximate induction strategies. In Chapter 5, I test how people learn causal structures by observing events unfolding in continuous time. I find that people are capable learners in this setting, successfully identifying the large majority of generative, preventative, and non-causal relationships but making certain attribution errors. I build a model that approximates normative inference via a simulation and local summary statistics scheme and shows it better captures participants' judgment patterns than the normative account, indicating that people make continuous-time causal induction based on structurally local computation using temporally local evidence.

Besides passive observations, real life also requires people to interact with causal systems, choosing actions or "interventions" in order to gather information about how a system works. In Chapter 6, I test how people actively learn about causal structure in continuous time, focusing on when and where they intervene in a causal system, and how this shapes their learning. I find

that people time and target their interventions to create simple yet informative causal dynamics. I propose a resource-rational account to explain how people balance expected information and inferential complexity when choosing interventions. I discuss how the continuous-time setting challenges existing computational accounts of active causal learning, and argue that metacognitive awareness of one's inferential limitations plays a critical role for successful learning in the wild.

Genuine causal influences can take complex forms and our measurements of them are inevitably incomplete: Some effects might occur instantly and dissipate rapidly, but others might peak later grow or compound over days or years. As such, uncertainty exists as to when is a good moment to end the observation. In Chapter 7, I test what people infer from a limited period of observation after an intervention. I find that under certain conditions people extrapolate what might happen in the future and rely on these extrapolations to make inferences, even drawing the opposite causal conclusion to people who believe the causal influence is spent by the end of the observation. This reveals how functional learning and generalization play a role in how people utilize temporal information to both learn and apply causal models to augment their understanding of the world.

Finally, in Chapter 8, I pull together the empirical and theoretical insights learned and propose a roadmap for future research on learning causality in time.

Chapter 2

Theoretical frameworks of causal induction

> "That the sun will not rise tomorrow implies no more contradiction than that it will rise."
>
> *David Hume*

People are intrinsically motivated to learn how our world works. However, as David Hume says, the reality is inherently uncertain. We can often only learn by collecting limited and noisy evidence and then inferring general rules to guide our prediction and control in the future. Causal induction theories are concerned with how people infer the nature of the causal relationships between entities or phenomena on the basis of information they have obtained. In this chapter, I will review the theoretical evolution and important empirical findings in human causal induction, by walking through several quantitative theories that have attempted to describe human causal induction over the past five decades. Each theory builds on earlier approaches, with later theories addressing empirical findings that remained unsolved by the preceding theories. As we trace the historical trajectory of causal cognition theories, the significance of the ideas of *theory-based cognition* and *bounded rationality* will become evident, which will serve as two pivotal theoretical guidelines for this thesis.

Towards the end of this chapter, I will discuss the challenges confronted by previous models. These challenges include the inability to (1) predict causal direction, (2) determine the time window for observation, (3) contend with evidence from repeated measures, and (4) learn cyclic structures. All of these challenges converge upon a central theme: the important role of *time* in

causal inference. Later chapters will test these challenges empirically, with each chapter strategically highlighting different aspects, and develop novel causal learning models that can leverage temporal information to effectively address these challenges.

2.1 Three atemporal causal learning rules

2.1.1 Rescorla-Wagner rule

Empirical research about the recognition of relationships between events can be traced back to Behaviourism, one of the most influential approaches in 20^{th} century psychology. Behaviouristic scientists focus on how human and non-human animals would associate binary variables based on the experience of statistical contingency as well as spatiotemporal contiguity (Pavlov, 1928; Skinner, 1938). Their experimental paradigms often include training stages that display associated or disassociated evidence and testing stages that examine associative strengths by measure subjects' behaviour patterns (see Shanks, 1995; Pearce & Bouton, 2001, for review).

Rescorla & Wagner (1972) formalise experienced association strengths on the basis of co-occurrence evidence (i.e. subjects experienced evidence trial by trial, and in each trial different stimuli may be present or absent) with a simple mechanical rule – the RW rule. Intuitively, it suggests that learning occurs to the extent that a learner feels "surprised" about a new observation. For instance, if a learner believes there is no relationship between A and E but then observes A and E's co-occurrence (which is denoted as $A = E = 1$), they should slightly increase their association between A and E. However, with repeated exposures to $A = E = 1$, the association starts to asymptote because the evidence becomes less surprising to the learner. The RW rule can also be applied to situations when multiple events are associated with a target event (Saavedra, 1975). If the learner already believes A and E are associated, the later observation of $A = B = E = 1$ will be not surprising and therefore cannot increase the association between B and E, which is called the *forward blocking* phenomenon (Kamin, 1967; Le Pelley et al., 2017).

More specifically, the RW rule depicts how beliefs change dynamically as trial-based information flows in:

$$V_{C_1}^t = V_{C_1}^{t-1} + \Delta V_{C_1}^t \tag{2.1}$$

$$\Delta V_{C_1}^t = \alpha\beta(\lambda - \sum_{C \in \mathbb{C}_t} V_C^{t-1}) \tag{2.2}$$

Equation 2.1 states that the associative strength at the current trial t depends on the original strength plus the change due to trial t. Equation 2.2 specifies that in a given trial, the strength for particular cause C_1 is updated according to both whether the effect E co-occurs with C_1 ($\lambda = 1$) or not ($\lambda = 0$) and the existing association strength based on how many causes occur in the

Table 2.1: The format of two by two causal tabular data.

	Cause=1	Cause=0
Effect=1	a	b
Effect=0	c	d

current trial as well as their predictive strengths respectively. Two fixed learning rate parameters α and β are added which depend on the salience of C_1 or outcomes.

The RW rule assumes that subjects learn a network of associations rather than setting out to learn a causal model of the world. Despite the fact that RW has proven a successful predictor for many aspects of human and non-human animals' behaviour (Allan, 1993), it fails to predict several phenomena especially in human subjects. First, although *forward blocking* is often shown in non-human animals, it is relatively weak or even fails to observe in the human learning process (Kamin, 1967; Shanks, 1985; Cheng & Lu, 2017; Le Pelley et al., 2017). Second, as a contrast to forward blocking, the phenomenon called *backward blocking* describes that when people experience co-occurrences of Cause A, Cause B, and Effect E (i.e. $A = 1$, $B = 1$, $E = 1$), and are then trained on co-occurrences of only A and E (i.e. $A = 1$, $E = 1$), their causal strength judgment of B will decrease (Le Pelley & McLaren, 2001; Shanks, 1985; Wasserman & Berglan, 1998). However, since there is no information about B at the second stage, the RW rule does not predict this updating of B's causal strength. Third, the ability to infer *unobserved causes* found in both humans (e.g. Lipp & Vaitl, 1992) and non-human animals (e.g. Hall & Honey, 1989) are at odds with the RW rule that only considers observed variables (see Gershman et al., 2010; Redish et al., 2007, for review and computational explainations). All three phenomena suggest that causal learning may not be simple reflections of associations but are additionally sensitive to one's mental representation of the underlying causal structure.

2.1.2 Delta-P

Historically, human cognition researchers focus on how humans make causal inferences from descriptions such as whether a certain fertiliser will cause plants to bloom. In these studies, human participants could be asked to experience event associations through trials, but also could just read the summarised statistical information. The information of two binary variables is usually presented as *2-by-2 tables* (see Table 2.1). As shown in Equation 2.3, the Delta-P rule (Allan, 1980; Jenkins & Ward, 1965) assumes that people infer causal strength by comparing cases that effect occurs with the cause present, with cases that effect occurs with the cause absent. Generative causal judgments equal to ΔP and preventative causal judgments equal to $-\Delta P$:

$$\Delta P = P(E = 1|C = 1) - P(E = 1|C = 0) = \frac{a}{a + c} - \frac{b}{b + d} \tag{2.3}$$

As with the RW rule, Delta-P is also an associative quantity that does not address any mental causal representation. It provides a solution for prevailing scenarios in human life and scientific discovery. Indeed, since then, inferring from contingency tables has received a lot of attention in the field of causal reasoning research. Delta-P performs better than many other calculations in predicting causal strength judgments (Allan & Jenkins, 1983), but it is insensitive to the "density bias" found in humans (Allan & Jenkins, 1983; Baker et al., 1989; Buehner et al., 2003; Shanks & Dickinson, 1991): If a set of scenarios are constructed in which ΔP is fixed while other aspects of the contingencies are varied, human generative causal strength judgments do not remain constant. Specifically, they tend to increase as $b/(b+d)$ increases, and preventative judgments decrease as $b/(b+d)$ increases. The causal power theory described below successfully addresses this problem.

2.1.3 Causal power

As Delta-P, the causal power theory (aka. Power PC, see Cheng, 1997; Cheng & Lu, 2017, for review) also aims to extrapolate causal strength between binary variables with evidence that can be formed as a 2-by-2 table (Table 2.1). Compared to associative theories, Power PC demonstrates four assumptions about human causal reasoning:

- There is an unobserved cause B that can produce the effect E but not prevent it.
- The evaluated cause C and the unobserved cause B influence E independently.
- The power of a cause is independent of the frequency of occurrence of the cause.
- E does not occur unless it is caused.

The core feature of Power PC is assuming an unobserved hidden cause that accounts for the effect's presence when the observed cause is absent. Accordingly, when C is a generative cause, the effect could be caused by either C or B, and therefore the probability of observing E follows a *noisy-OR* function:

$$P(E=1|w_b, q_c) = q_c \cdot c + w_b - q_c \cdot c \cdot w_b \tag{2.4}$$

The $c \in \{0,1\}$ represents the absence or presence of C, q_c represents the causal strength of c, and w_b represents $q_b \cdot b$. Given that $P(E=1|w_b, q_c) = q_c + w_b - q_c \cdot w_b$ when C is present and $P(E=1|w_b, q_c) = w_b$ when C is absent, we can derive Equation 2.5 as the calculation of causal strength q_c:

$$q_c = \frac{P(E=1|C=1) - P(E=1|C=0)}{1 - P(E=1|C=0)} \tag{2.5}$$

If C is a preventative cause, the effect could be caused by B but then possibly presented by C. The probability of observing E follows a *noisy-AND-NOT* function in Equation 2.6. Accordingly,

the preventative causal strength can be represented as Equation 2.7.

$$P(E = 1|w_b, q_c) = w_b \cdot (1 - q_c \cdot c) \tag{2.6}$$

$$q_c = \frac{P(E = 1|C = 0) - P(E = 1|C = 1)}{P(E = 1|C = 0)} \tag{2.7}$$

We could find that Power PC can be seen as Delta-P adjusted by the effect's base rate ($b/(b+d)$ in Table 2.1). It considers the proportion of cases that E is already present or absent and therefore C, as a generative or preventative cause, would have no chance to affect E. Power PC successfully predicts the phenomenon that ratings of generative influences become higher when base rates are high (which implies the cause would have succeeded a greater proportion if it had more space to operate), and stronger preventative influences when base rates are low (Buehner et al., 2003; Cheng, 1997). It also predicts that when the base rate equals to 1 in generative causal judgments (or 0 in preventative causal judgments), people would consider evidence to be uninformative to infer causal relationships (Wu & Cheng, 1999). However, there are still findings on human causal judgments that Power PC fails to explain. For example the "frequency illusion" phenomenon demonstrates that when $\Delta P = 0$, people's causal judgments would become stronger as the number of observed data points increases (Griffiths & Tenenbaum, 2005), which is inconsistent with Power PC that predicts a constant judgment at zero. The theory-based causal induction approach below successfully addresses this problem.

2.2 Theory-based causal induction

From the development of three atemporal causal learning rules, we can see the significance of mental model representation is gradually recognized. Incorporating reasoning about background causes in Power PC implies that local causal judgments are not made in isolation, but rather take surrounding *structure* into account. Indeed, the real life situations could also often include complex structures. For example, it may be reasonable to presume that depression causes insomnia, and insomnia causes anxiety (i.e. a chain structure, see Figure 2.1 for the illustration of structures with different names), or alternatively that depression causes insomnia and anxiety independently (i.e. a fork structure). People demonstrate the ability to incorporate different mantel models when make inferences. In Waldmann & Holyoak (1992), participants first experienced the co-occurrence of AB and AC to establish a perfect, deterministic connection between the two pairs. They were later presented with situations where the status of A was masked, and they had to predict B's status given C. Participants who were told that A was a virus and B and C were symptoms predicted the presence of B when C was present, given that they were common effects of A. However, participants who were told that A was an emotion and B and C were appearance features that could affect emotions did not predict B's presence when C is present, as B and

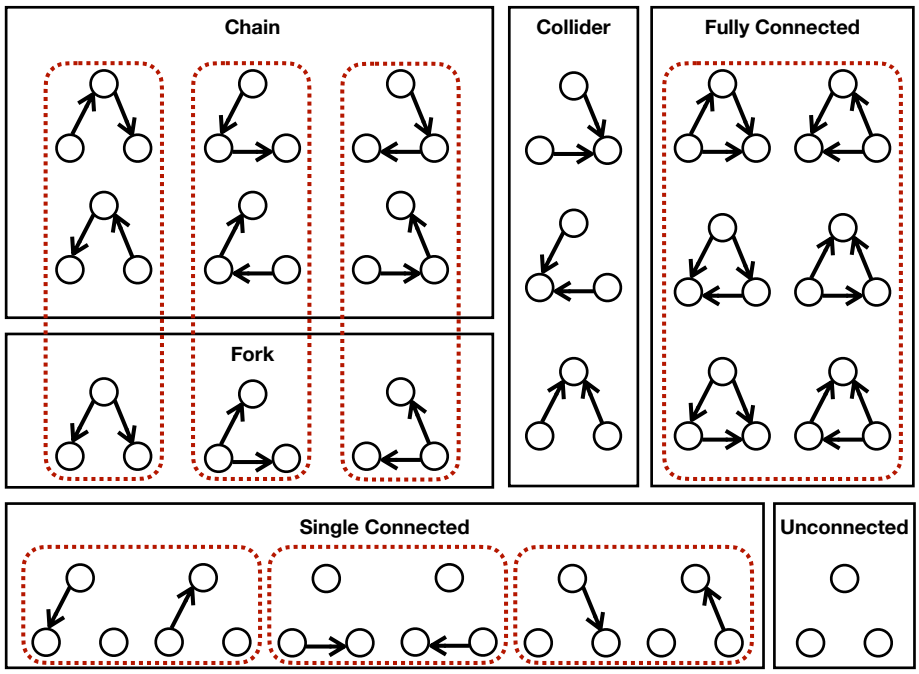


Figure 2.1: Directed acyclic graphs (DAGs) on three variables. Red boxes indicate Markov equivalence structures. Fork structures are also called common-cause structures. Collider structures are also called common-effect structures.

C were seen as separate causes of A. This indicates that people form *directional* causal models based on instructions and utilize them to predict outcomes, whereas association theory does not distinguish between the two situations (see also Waldmann, 2000; Waldmann & Hagmayer, 2005).

2.2.1 Ontology, plausible relations, functional form

How do people make model-based causal judgments? Griffiths & Tenenbaum (2009) highlights three critical components of the rational causal induction process: An ontology that outlines the entities under investigation and their properties, a set of plausible relations that suggest how entities may be connected, and the functional form that determines how causes influence their effects under each type of relations.

Using the graphical illustrations and Bayes' rule (Sloman, 2005; Pearl, 2000; Griffiths & Tenenbaum, 2009), we can interpret this process as requiring the identification of nodes (i.e., entities) within the causal structure of interest, the construction of a hypothesis space that includes links (i.e., connections) between nodes, and the specification of likelihood functions to evaluate the proposed structures (see Box 2.1 at the end of this chapter for more introduction of Bayes' rule).

The learner updates their prior belief over each structure s in the hypothesis space $P(s)$ with a likelihood function $P(\mathbf{d}|s;\mathbf{w})$ to get the posterior distribution $P(s|\mathbf{d};\mathbf{w})$, given data $\mathbf{d}$ and a set of parameters $\mathbf{w}$:

$$P(s|\mathbf{d};\mathbf{w}) \propto P(\mathbf{d}|s;\mathbf{w}) \cdot P(s) \tag{2.8}$$

This rule applies to both atemporal and temporal datasets. I next introduce causal Bayesian networks as a rational approach for the atemporal setting and formalize a rational approach for processing temporal data in Chapter 4.

2.2.2 Causal Bayesian networks

Causal Bayesian Networks (CBNs; Pearl, 2000) provide a principled approach to formalize complex causal relationships among multiple variables. It was developed for modeling large datasets in computer science contexts, but subsequent research shows that laypeople have a commensurate intuitive understanding of complex causal structures and both learn and make inferences that broadly reflect the predictions of CBNs theory (Steyvers et al., 2003; Griffiths & Tenenbaum, 2005, 2009).

CBNs are kinds of probabilistic directed acyclic graphical models. They represent variables as nodes and causal relationships as arrows and are defined by the "Markov assumption" that once all the direct causes of a variable X are controlled for, X must be statistically independent of other variables in the causal network that are not its direct or indirect effects. CBNs provide ideas for causal structure learning from both qualitative and quantitative aspects. The qualitative aspect is that causality is represented as a network of direct dependence between variables. For example, if we want to confirm that depression causes insomnia and anxiety independently (insomnia $\leftarrow$ depression $\rightarrow$ anxiety), there should be statistical dependence between insomnia and depression, and depression and anxiety, but insomnia and anxiety should be irrelevant once the state of depression is known.

Quantitatively, it follows the Bayes' rule that one can incorporate a prior belief with the likelihood of newly observed data to form an updated "posterior" belief. In this case, this is a posterior over all hypothetical causal structures and then choose the most likely causal structure. The likelihood $P(\mathbf{d}|s;\mathbf{w})$ is calculated according to Equation 2.9, where $\mathbf{d}\{e_1, ..., e_i\}$ represents the effect in each trial. The $genPa(e_i)$ and $prePa(e_i)$ are datasets of generative or preventative parent nodes associated with e_i. The calculation of $P(e_i|Pa(e_i))$ refers to noisy-OR and noisy-AND-NOT functions (Equation 2.10), which is similar to Power PC despite that now one effect can be influenced by multiple causes (and the base rate could be regarded as one cause in $genPa(e_i)$). The generative and preventative causal power parameters $\mathbf{w}\{q_g, q_p\}$ could be learned via instructions

or jointly learned with causal structures (Griffiths & Tenenbaum, 2005, 2009).

$$P(\mathbf{d}|s;\mathbf{w}) = \prod_i P(e_i|genPa(e_i), prePa(e_i)) \tag{2.9}$$

$$P(e_i|genPa(e_i), prePa(e_i);\mathbf{w}) = [1 - \prod_{g\in genPa(e_i)} (1 - q_g \cdot g)] \prod_{p\in prePa(e_i)} (1 - q_p \cdot p) \tag{2.10}$$

CBNs provide a comprehensive theoretical framework that can handle a wide range of human causal reasoning questions across different domains (see Rottman, 2017, for review). Compared to Power PC, CBNs not only can reflect the inner representation of unobserved hidden causes, but also incorporate Bayesian priors to capture potential human inductive biases. Griffiths & Tenenbaum (2005) encode two-variable causal strengths usually demonstrated under Power PC into CBNs. Griffiths & Tenenbaum (2005) explain the frequency illusion by assuming that what people actually do is to distinguish between two causal hypotheses: a *Graph 1* where both unobserved background causes and the target cause are linked to the effect, and *Graph 0* where only the unobserved background causes are linked to the effect. Under limited observed data, people are uncertain about both graphs, and therefore the causal judgment – the normalized probability of Graph 1 – is larger than that after people gather enough evidence to support Graph 0. Essentially, they argue that what people infer in these settings is not the strength of association between cause-effect pairs, but the *probability* that a causal link exists. Thus, they sometimes replace the term "causal strength" with "causal support" to describe the belief about the link between a putative cause and effect.

2.3 Bounded rationality

Marr's three levels of analysis have been widely acknowledged in cognitive science research (Marr, 1982). He argues that an information processing system can be analyzed at three levels: (1) The computational level (also called the normative level) that explains the problem that the system is going to solve and how an ideal or rational agent would solve the problem; (2) the algorithmic level (also called the process level) that explains how a system is solving the problem and what algorithm is implementing that solution; and (3) the implementation level that explains how the algorithm is physically implemented.

Now let us rethink classical causal theories under Marr's framework. Both the RW rule (Rescorla & Wagner, 1972) and Delta-P (Allan, 1980) aim to describe empirical behavior so they are often classified as process-level models. Power PC (Cheng, 1997), as well as causal Bayesian networks (Griffiths & Tenenbaum, 2005, 2009) can be regarded as both normative and process accounts for atemporal causal learning given that the authors not only provide proofs of why their solutions are optimal, but also show that people's performances are best fitted by their models.

Tauber et al. (2017) point out that Bayesian models have a broad interpretive power (i.e. the ability to capture a wide range of phenomena) since there can be a large number of combinations between priors and likelihood functions. By using the optimal combination, Bayesian models can provide optimal solutions. But in other times, it can also serve as a language to describe learners' mental representations, beliefs under these representations, and learning rules for belief revisions, by specifying correspondent bounded hypothesis spaces, priors, and likelihoods, respectively.

Although aggregate results can often be in line with normative Bayesian models (Goodman et al., 2011; Gopnik & Tenenbaum, 2007), any single individuals' judgments are typically much noisier and more idiosyncratic (Vul et al., 2014; Tauber et al., 2017). In order to utilize Bayesian models as algorithmic-level models, researchers need to examine their predictions for individual results, which is often understated in previous causal learning studies (Griffiths & Tenenbaum, 2005, 2009; Pacer & Griffiths, 2012, 2015). Moreover, researchers sometimes consider a model could be both normative and process by admitting that learning tasks are easy (i.e. less cognitive demanding) and hence human can perform optimally under their cognitive capacity. However, it is also possible that people are using approximations that are indistinguishable from optimal calculations because both solutions can provide optimal answers to an easy problem. Accordingly, a more practical way to study causal learning and find process models would be setting up moderately difficult tasks that can potentially separate the prediction of normative- and process-level solutions and examining both aggregate and individual results (Van Rooij et al., 2019) — that is the approach adopted by this thesis.

2.3.1 Rational analysis

How can we find process-level models that can better explain *human causal learning*, especially in terms of learning complex causal structures? Modeling human cognition, a black-box system that is much more intelligent than anything we have ever created ourselves (Lieder & Griffiths, 2020), could be a challenging task. Given an observed behavior pattern, researchers can propose an infinite number of models that are capable of explaining the data, but for many times at most one of them is correct. Therefore, rather than "discovering" the human cognitive processes, it is better to describe the job of cognitive researchers as "approximating" the truth of human cognition. *Rational Analysis* (Anderson, 1990) provides guidance for this process. It requires researchers to 1) develop an optimal model formulated as a problem and its solution, with minimal assumptions about computational limitations; 2) examine the empirical literature to see if the predictions of the behavioral function are confirmed; 3) if the predictions are off, then refine the model to incorporate more constraints and better capture the data.

Table 2.2: The number of possible generative structures for 1-6 variables.

Variables	Acyclic only	Acyclic and Cyclic
1	1	1
2	3	4
3	25	64
4	543	4096
5	29281	1048576
6	3781503	1073741824

2.3.2 Three process-level causal learning models

Based on the rational analysis procedure, researchers have developed process-level algorithms to enhance our understanding of how humans learn causal structures in comparison to normative CBNs. In this context, I will introduce three such algorithms, all tailored to address problems related to causal systems comprised of binary variables. Each of these algorithms demonstrates the capacity to approximate the results obtained from CBNs over an extended period by either observing a greater amount of data (win stay lose sample) or executing a more extensive sequence of sampling (Neurath's ship and mutation sampler). Additionally, these algorithms incorporate constraints on computational resources to provide a more accurate representation of bounded human cognition.

Win stay lose sample

Bonawitz et al. (2014) demonstrate that causal learners employ a particle-based approximation called *win-stay-lose-sample* (WSLS) for reasoning. This approach involves learners maintaining a single sample belief from a hypothesis space, which is subsequently resampled when it fails to generate evidence consistent with the observed data. In other words, as the current hypothesis becomes less capable of explaining the most recent data, the probability of resampling increases. The advantage of WSLS lies in the fact that, when confronted with a new data point, the learner only needs to evaluate its likelihood under the currently focused hypothesis, rather than examining it under all possible hypotheses.

In comparison to an ideal Bayesian learner, the use of WSLS introduces a phenomenon known as "stickiness" – a tendency for the learner to favor the currently held hypothesis over alternative hypotheses. This preference aligns with the conservatism observed in broader human cognition (Bramley et al., 2015; Edwards, 1968; Phillips & Edwards, 1966). WSLS has successfully explained this sequential dependence in online causal judgments made by both adults and children.

Neurath's ship

Although WSLS suggests that humans may only evaluate each data point under a limited number of hypotheses, it still requires the learner to maintain a belief distribution encompassing all possible hypotheses for the resampling process. This requirement can be met in certain scenarios, such as the one described by Bonawitz et al. (2014), where only 16 hypotheses are employed in the deterministic setting and only three hypotheses in the probabilistic setting. For example, in their deterministic task, participants were asked to determine which type of block could cause which type of block to light up (e.g. red→yellow, yellow→yellow, yellow→red, red→red), where the 16 hypotheses can be further decomposed into four independent rules.

However, real-life causal learning problems could involve more intricate causal structures. For instance, even when considering only acyclic structures, the number of possible generative causal structures (excluding preventative structures) for just three variables would amount to 25. If cyclic structures are also taken into account, this number increases to 64 (refer to Section 2.4.4 for a detailed explanation of the distinction between acyclic and cyclic structures). As illustrated in Table 2.2, the number of structures grows quickly and becomes intractable as the number of variables increases. Furthermore, the variables within these structures can be interconnected in more complex ways, such as forming chains (e.g., $A \rightarrow B \rightarrow C \rightarrow D$). Interestingly, even though people often cannot perfectly discover the underlying causal structures, they perform reasonably well in tasks involving three or four variables, both in non-temporal settings (Bramley et al., 2015; Bramley, Dayan, et al., 2017) and temporal settings, as will be demonstrated in the subsequent chapters of this thesis.

Bramley, Dayan, et al. (2017) show that people's online causal learning is similar to the *Neurath's ship* metaphor in philosophy of science, where learners only adjust their hypothesis partially each time (i.e. local updating), and do so using limited recent evidence (i.e. local information). Specifically, the Neurath's ship model assumes that learners hold a single hypothesis regarding the causal structure in their mind. When confronted with new evidence, they iteratively search for local improvements to the currently focused causal structure by adding or subtracting links step by step, or reorienting existing links. The direction of these adjustments is determined using a Markov chain Monte Carlo sampling process (Goudie & Mukherjee, 2011).

Markov chain Monte Carlo (MCMC) is a sequential sampling technique employed to approximate probability distributions in cases where exact calculations are computationally infeasible. For the recent decades, MCMC has been adopted in cognitive studies as a tool to approximate posterior distributions (Lieder, Griffiths, Huys, & Goodman, 2018; Lieder, Griffiths, & Hsu, 2018; Dasgupta et al., 2017; Bramley, Dayan, et al., 2017; Davis & Rehder, 2020). It allows researchers to investigate how individuals approximate the posterior distribution when analytical computations are intractable or impossible.

Metropolis–Hastings (MH) sampling and Gibbs sampling are two widely recognized types of MCMC algorithms. MH sampling proposes a set of new parameter values, deciding whether to move to the new set or stay with the original set based on the relative probabilities between the new and old sets of values. On the other hand, Gibbs sampling, a special case of MH algorithms, samples from conditional probability distributions and always accepts the samples. That is, it samples a new value for one parameter while keeping the previous state unchanged. This process continues iteratively for each parameter in the model until the value of each parameter converges. In the context of the Neurath's ship algorithm and its local changes to causal structure hypotheses, it aligns with the principles of Gibbs sampling (Bramley, Dayan, et al., 2017). By sequentially sampling for a large number of iterations, the values in MCMC algorithms often converge to the true underlying values or states. However, to account for cognitive resource limitations, it is common to assume that individuals can only sample a limited number of times. In the case of the Neurath's ship algorithm, it assumes that the local search will terminate after a fixed number of steps (e.g. 50 in their experiments). After this point, the memory of evidence will be cleared if the held structure has been updated to a different one. Subsequent updates will solely rely on later pieces of new evidence, aligning with the notion of *local (recent) evidence* (Bramley, Dayan, et al., 2017).

The algorithm is referred to as the "Neurath's ship" due to its resemblance to the corresponding philosophical metaphor (Quine, 1960), which suggests that, similar to sailing a ship, people rely on their existing theories or hypotheses to navigate through uncertain waters. We continuously improve and refine our theories while being unable to retreat to a dry-dock to contemplate all possible alternatives and make significant changes. The Neurath's ship model demonstrates a better overall fit to human causal judgments compared to WSLS. This finding indicates that, instead of completely resampling a new structure from the hypothesis space, individuals are more inclined to refine their existing hypothesis through local adaptations and adjustments.

Mutation sampler

WSLS and Neurath's ship models provide insights into human causal judgments by considering how the *hypothesis space* can be narrowed down. However, they still assume that the reasoner uses the normative Bayesian approach to calculate how likely it is to see the evidence given a hypothesis (see Equation 2.9 and Equation 2.10). In contrast, the mutation sampler, as the third algorithm, highlights another significant perspective: How the *likelihood calculations* could be more computationally economical.

The mutation sampler algorithm (Davis & Rehder, 2020) proposes that causal inferences are made based on a set of samples representing possible states of the causal system. Rather than relying on exact probabilistic calculations, this approach assumes that the likelihood of seeing a state (e.g. $\{A = 0, B = 1, C = 1\}$) originating from a causal structure (e.g. $A \rightarrow C \leftarrow B$) depends

on the number of that state sampled from that particular causal structure. The sampling process is sequential and follows the MH sampling approach: At each step, a new state is "mutated" from the previous state by modifying the value of one variable in the structure. A key assumption of this algorithm pertains to the initial state. It assumes that the initial state is one of two possibilities: either all causes (variables without endogenous parents) are set to 0, or they are set to 1, and the effects work by assuming all causal links in the system are deterministic. For example, in a fork structure such as $A \rightarrow C \leftarrow B$, the initial state would be either $\{A = 0, B = 0, C = 0\}$ or $\{A = 1, B = 1, C = 1\}$. Similar to the Neurath's ship, the mutation sampler imposes constraints on the number of samples to account for cognitive limitations. As a result, there may be a higher number of initial states compared to other states in the final distribution, reflecting the limited number of samples that individuals can generate.

The mutation sampler, through its departure from the normative causal Bayesian network account, offers explanations for various classic fallacies observed in atemporal settings. These deviations align with how humans often diverge from the normative account. To further investigate how individuals simulate mental samples, Davis & Rehder (2020) conducted an experiment in which participants were asked to distribute coins among the states they sequentially sampled. The observed sampling distributions exhibited greater alignment with the mutation sampler rather than the normative sampler. These findings suggest that individuals may possess a "prototype" representation of causal structures and only briefly explore alternative possibilities before settling on a particular hypothesis (Davis & Rehder, 2020).

2.3.3 Resource rationality

Lieder & Griffiths (2020) expanded the rational analysis framework by emphasizing the concept of bounded rationality. They proposed that human behavior is not in opposition to rationality but rather represents a bounded version of rationality, as initially described by Simon (1982). According to this perspective, individuals rely on their limited cognitive capacity to solve problems by making optimal trade-offs between the cost of computational resources and the utility of achieving a more accurate approximation of the correct solution. In this framework, known as "resource-rational analysis", researchers develop process models that take into account the cognitive constraints faced by individuals, including but not limited to elementary operations, processing speed, and working memory capacity.

A crucial step in resource-rational analysis is to *quantify* the cognitive cost associated with different actions or strategies. For example, when using the MCMC algorithm, researchers can assume that each sample incurs a certain cost, which is a function of the number of samples taken. The reward r of a decision is determined by the reward defined by the task at hand. Traditionally, external rewards have been considered, but recent research has also explored the quantification of internal rewards, such as curiosity or intrinsic motivation (Brändle et al., 2023). By making a

bounded optimal action a^*, learners aim to maximize their overall reward r while minimizing the cognitive costs c:

$$a^* = \underset{a \in \mathbf{A}}{\text{argmax}}(r_a - c_a) \tag{2.11}$$

Within the framework of resource-rational analysis, many human behaviors that were previously considered irrational have been reinterpreted as rational adaptations, taking into account the constraints of human cognitive resources and the specific ecological context. This paradigm shift has led to a better understanding of human decision-making processes across different domains, including (1) choosing between machines with different reward distributions in decision tasks (Binz et al., 2022; Lieder & Griffiths, 2017); (2) answering open-ended questions that may involve anchoring-and-adjustment process, such as estimating the frozen degree of vodka (Dasgupta et al., 2020; Lieder, Griffiths, Huys, & Goodman, 2018); and (3) planning sequential actions (Callaway et al., 2022). In Chapter 6 of this thesis, the same resource-rational analysis framework will be applied to explain another category of decision-making: intervention decisions aimed at learning causal structures. By considering the limited cognitive resources people possess and the specific challenges posed by causal learning, this framework promises to shed light on the rational adaptations underlying intervention decision-making processes.

It is important to note that resource rationality does not provide a direct answer regarding the specific algorithms implemented by individuals. For instance, in the context of planning and causal intervention decisions, calculating the exact computational cost and normative rewards for each available option may require more cognitive resources than simply computing the normative rewards and making a decision based on rewards alone (Callaway et al., 2022). Therefore, compared to treating it directly as a process-level model, it is more reasonable to recognize that resource rationality allows us to explore the extent to which individuals consider cognitive costs in their decision-making.

2.4 Limitations of existing causal learning models

So far, I have presented an overview of classical causal theories (RW, Delta-P, Power PC, CBNs), along with recent process-level models (WSLS, Neurath's ship, mutation sampler). These frameworks offer quantitative explanations for how humans make causal inferences. However, despite their contributions, several challenges remain unresolved within these influential frameworks. These challenges, which I will outline below, primarily stem from the neglect of temporal information.

2.4.1 Causal direction

All causal models reviewed above focus on answering the question of whether, or to what extent, variables are *associated* with one another, but they are not able to discern the *direction*: which variable is the cause and which is the effect. Since the RW rule and Delta-P are cast as associative learning theories, they do not aim to solve this causal direction problem. In studies of Power PC, the causal direction is often indicated in cover stories that people can easily understand: e.g. that drugs affect symptoms or chemical substances affect bacteria growths but not vice versa (Buehner et al., 2003; Wu & Cheng, 1999), based on prior knowledge obtained in daily life rather than mechanical thinking about the current evidence (Lagnado et al., 2007).

CBNs have more flexible hypothetical causal structures, while as shown in Figure 2.1, there are many structure groups that cannot be distinguished from each other on the basis of observed data because their dependencies are all equivalent. These are known as called Markov equivalent structures. For example, in casual structures $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \leftarrow Z$, and $X \leftarrow Y \rightarrow Z$, the dependence patterns are all that X and Y are always correlated, Y and Z are always correlated, and X and Z are unconditionally correlated but become uncorrelated once Y is controlled for. Unlike with traditional Bayesian networks, for CBNs Markov equivalent structures can be distinguished by introducing intervention data. *Intervention* means to manipulate one or more of the variables in the model to one possible value, so that the value of these variables will be fixed and no longer depend on its parents. If you intervene on Y, and find Y and Z are correlated, but X and Y, X and Z are independent, then $X \rightarrow Y \rightarrow Z$ is the correct structure (Pearl, 2000). Intervening opportunities are not always readily available, yet in our everyday lives, we can easily distinguish causes from effects. How do we accomplish this? The answer may lie in the consideration of temporal information, which none of the aforementioned theories have explicitly incorporated or addressed. In Chapter 3, I will dig into the existing literature on the significance of temporal information in the realm of causal reasoning.

2.4.2 The observation window

All of the models mentioned in this chapter deal with atemporal contingency data, where one widely used paradigm involves presenting participants with pairs of events in independent samples. Cover stories in scientific fields such as biology (Buehner et al., 2003; Lu et al., 2008), physics (Lagnado & Sloman, 2004; Coenen et al., 2015), and psychology (Rottman & Keil, 2012) are used, since the data structure is similar to the data that scientists collect under laboratory experimentation: In order to obtain convincing and generalizable knowledge, scientists are required to collect independent evidence of sufficient sample size (Hattori & Oaksford, 2007). A minimal example might involve pairs of patient outcomes (e.g. sick or not) under different treatment assignments (e.g. vaccinated or unvaccinated) and ask them to judge whether or to what extent the treatment affected the outcome (Buehner et al., 2003; Stephan, Placì, & Waldmann, 2021).

While these settings put timing considerations to one side, they do not eliminate them. Fundamental questions remains as to how to determine an appropriate time window to measure outcomes, and how to ensure the observations are truly independent of one another. Without supporting knowledge about the relevant causal mechanisms, waiting too short a time before measuring an effect may not allow the influence to propagate or become apparent (e.g. the vaccine may not have taken effect yet), while waiting too long will tend to introduce confounds factors (e.g. the infection running its course, or the patient dying from natural causes). Equally, we need to determine the timing of interventions since some time-dependent factors (e.g. age) mediate the relationships between variables (Rottman, 2016). It appears that, to construct these simple contrasts, scientists are already using rich prior causal knowledge about the relevant mechanisms and their temporal properties in order to create the experimental protocol that allows abstraction to the level of contingency data. If so, it is important to understand how people acquire these temporal expectations in the first place. In Chapter 4, I will build a rational framework to show how the temporal expectations could be learned from causal learning process. In Chapter 7, I will show how important the observation window is in influencing causal judgments, which questions the previously simplified way to choose the observation window.

2.4.3 Independent vs. repeated-measure evidence

The other problem with atemporal contingency data is that, although it is common to see in the scientific discovery, it may systematically deviate from everyday life situations. Scientific samples typically consist of independent observations, a concept formally denoted as "independent and identically distributed" (i.i.d.). In contrast, lay people often experience multiple events of the same type occurring multiple times to a single individual. For example, if we want to learn how to use the TV remote control, we will probably try pressing buttons repeatedly rather than observing many independent televisions and controllers being used by others. If we want to know what activities can improve working efficiency, we probably try them ourselves on different days rather than asking hundreds of friends to change their schedules. What laypeople gather and infer upon is generally “repeated-measures” evidence rather than independent and identically distributed evidence.

On one hand, the amount of samples could be limited compared to scientific experiments, leading people to rely on other cues (e.g. time, prior knowledge) rather than the contingency principles (Lagnado et al., 2007). On the other hand, the temporal dimension can also complicate the information in each data point, requiring sophisticated approaches to process it efficiently. Empirical studies indicate that people are sensitive to time when inferring causal structures (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Hagmayer & Waldmann, 2002; Greville & Buehner, 2010; Buehner & McGregor, 2006; Shanks et al., 1989; Valentin et al., 2022), while it has not been explained systematically how temporal information is integrated in the inference process. It is

unclear whether we actively transform this rich real-time data into the tabular formats that classical theories assume. If any, what strategy do people use? If people do not convert real-time data into tabular data, what theory could they rely on to make inferences? I argue that to improve our understanding of everyday causal induction, research should focus on settings with finer-grained repeated-measures evidence unfolding in continuous time. In Chapter 5 and 6, I will build causal learning tasks that requires the learner to learn from events from the continuously and examine their performance across different conditions.

2.4.4 Cyclic structures

Certain causal structures are only feasible when considering the time dimension, such as cyclic structures. A causal mechanism is cyclic if it has at least one component whose descendants include itself (Pearl, 2000). This means that the components that form part of the cycle, or outputs from it, may occur in repeated alternating fashion (e.g. a bidirectional connection $A \leftrightarrow B$ could generate a sequence of events A,B,A,B,A, ...). Many causal processes in the natural world are cyclic (Malthus, 1872), and people frequently report causal beliefs that include cyclic relationships when allowed to do so in experiments (Kim & Ahn, 2002; Nikolic & Lagnado, 2015; Sloman et al., 1998; Rehder, 2017). However, most influential causal learning models, such as CBNs, do not account for cyclic structures, instead focusing on directed acyclic graphs (DAGs; Pearl, 2000; Rottman & Hastie, 2014; Griffiths & Tenenbaum, 2009). Some adjustments have been made in order to capture cyclic structures, such as the use of dynamic Bayesian networks (Rottman & Keil, 2012; Valentin et al., 2022; Dean & Kanazawa, 1989). However, these models impose significant constraints on the data formats, limiting the expression of temporal information to discrete time steps (i.e. t, $t+1$, $t+2$, ...), and allowing each type of event to occur only once at each time step. These limitations do not reflect the true nature of events, which can happen at any moment in a continuous timeline, with intervals of any length between them. In order to study the way in which people reason about cyclic structures, appropriate tools are needed. In Chapter 6, I will conduct behavioral experiments that requires the learner to learn cyclic structures, and demonstrate a model that can capture cyclic structures.

Box 2.1: Probabilistic inference – Bayes' rule. Bayes' rule provides a solution to how we can use evidence to revise our beliefs, i.e. the problem of induction. According to the property of conditional probability that $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$, we can get Equation 2.12, where h represents hypothesis and d represents data:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \tag{2.12}$$

In many cases, we do not update a single hypothesis. For example, if there is a weighed coin and we wish to know whether it is "biased towards heads" or "biased towards tails", we will update the degrees of these two hypotheses simultaneously, with one increase and another decrease. We can form $H = \{h1, h2, h3, \ldots\}$ as the set to include all possible hypothesis we consider, which is often called *hypothesis space*. Then we can revise Equation 2.12 as:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in H} P(d|h')P(h')} \tag{2.13}$$

For most cases, we do not need to know the absolute probability of our hypotheses given the data, but relatively which hypothesis wins, so we can ignore $\sum_{h' \in H} P(d|h)P(h)$ since it is constant:

$$P(h|d) \propto P(d|h)P(h) \tag{2.14}$$

Equation 2.14 is the most commonly used formula in induction problems. $P(h)$ is called *prior* (or inductive bias) that reflect people's degrees of different beliefs before observing data. $P(d|h)$ is called *likelihood* where we calculate the probability of observing all data in d if a hypothesis is true. $P(h|d)$ is called *posterior* that combines the prior and likelihood to know the revised degrees of each belief, where we can finally choose the most likely belief to be the answer of the induction problem.

The likelihood function, i.e. how we get $P(d|h)$, would be an important piece that researchers need to illustrate in their works since it depends on the specific task. For the prior distribution, researchers can simply define a uniform prior distribution that all hypotheses are treated as equal possible before looking at the data, whereas if it is suspected that learners have biased prior beliefs, researchers also need to carefully consider the prior setting in their models. Finally, when the posterior calculation is intractable due to large hypothesis spaces or complex likelihood functions, researchers need to use some algorithms (e.g. Monte Carlo methods) to approximate the posterior distribution (see Griffiths et al. (2010) for a short overview of probabilistic inference in human cognition and Oaksford & Chater (2007) for a detailed one).

Chapter 3

Time, dynamics and causality

> "Lives are lived day by day, one day at a time, from day to day, day after day, day in and day out."
>
> *Kenneth Craik*

OUR lives are made up of the days we experience. In everyday experience, causal relationships are inseparable from temporal information in that events happen at particular times. We expect electronic equipment to turn on almost instantly after we push but that drugs will take hours to kill pain and that what we learn today may even influence our decisions ten years later. In this chapter, I will review empirical findings of the role of temporal information in causal learning processes.

3.1 What is "time"

A fun fact about time is that everybody seems to know what time is, but nobody can easily explain it to an organism who never experienced it before (Buonomano, 2017). The philosophy of time is often discussed by physicists. The uniqueness of time can be better understood through comparison with the spatial dimensions(S. M. Carroll, 2008, 2022):

1. We progress through time at a constant rate of one second per second, inevitably. Each moment is dependent on the preceding one: In a continuous spatiotemporal system, we can use information from one moment to make predictions about what will happen next. Conversely, it is far less reliable to predict events in one location based on information from another location.

2. The past is fixed, while the future remains uncertain: We often feel a sense of control over the future, but we understand that our present actions cannot alter the past. In contrast, the properties of left and right in spatial dimensions are identical. For example, our ability to see extends equally in both left and right directions.

3. Intuitively, we may doubt the existence of other points in time that we can travel to, but we are certain that there is something happening in other places that we are not in.

These philosophical properties of time indicate many relations between time and causality. The first property could be seen as the foundation of the Humean assumption that causes precede and are temporally contiguous with their effects (Hume, 1740). This property is also the most relevant one to this thesis, as I will show below. The second property indicates how people may control the outside world: We make actions upon a system and expect that the future rather than the past would change (Davis et al., 2018; Hagmayer et al., 2010). The third property conceives the process of counterfactual thinking where we mentally ruminate what could have happened had we acted differently in our past and then feel the emotions of luck or regret, and counterfactual thinking is regarded as an essential process in actual causal attribution (Lagnado et al., 2013; Gerstenberg et al., 2021).

Over several decades, cognitive researchers have investigated the relationship between time and causality. This inquiry has sought to understand both how temporal information serves as a cue towards causality and how causal beliefs shape one's perception of time. In the following sections, I illustrate different perspectives on time that researchers have diligently examined. These perspectives can be effectively categorized into two primary dimensions: delay information and order information.

3.2 Delay

3.2.1 Delay expectation

People tend to make stronger causal attributions when the delay between a putative cause and effect is consistent with prior expectations or mechanistic understanding (Gong & Bramley, 2023a; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Buehner & McGregor, 2006; Hagmayer & Waldmann, 2002; Stephan et al., 2020). For example, Buehner & McGregor (2006) found that participants assigned higher causal judgments to the insertion of a ball that turned on a light on a physical apparatus when the light came on after a few seconds, rather than instantly, if they were aware that it took time for the ball to roll through the apparatus and reach the light switch (see also Buehner & May, 2004). Similarly, Mendelson & Shultz (1976) test 4-7-year-old children on a machine that can make a bell ring if a marble is dropped into one of two holes. When children have little knowledge of the machine, they attribute it to the hole for which the bell rings

immediately after dropping a marble into it. However, after they observe the inner mechanism and realize marbles take time to go through a long tube, they choose the hole where the delay is long between marble dropping and bell ringing (see also Schlottmann et al., 2013).

Hagmayer & Waldmann (2002) found participants judged whether an insecticide prevents mosquitoes by comparing prevalence of mosquitoes in fields with and without the insecticide, but judged whether planting flowers prevents mosquitoes based on whether the prevalence of mosquitoes was affected the year after the flowers were planted, presumably expecting that flowers would take longer to influence the insect population than insecticide. All these studies highlight the role of expectation in relationships between causal latency and causal judgment: we select the cause that consists of what we expect how long the delay should be in the current context.

3.2.2 Short delay

People tend to make stronger causal attributions for short temporal delays than long ones, when mechanistic priors are not specifically conveyed (Shanks et al., 1989; Shanks & Dickinson, 1991; Greville & Buehner, 2010; Lagnado & Sloman, 2006; Buehner & McGregor, 2006; Greville & Buehner, 2007). Shanks et al. (1989) showed that people are more likely to learn a causal relationship between the action and the outcome when the outcome follows the action within two seconds. When the delay is longer, people are less likely to expect the action to bring about the outcome (see also Shanks & Dickinson, 1991; Greville & Buehner, 2010). This is consistent with the contiguity effect in humans' and animals' association learning, which shows that the association formed between two events decreases as the delay increases (Tarpy & Sawabini, 1974; Garcia et al., 1966).

At the same time, causal beliefs can, in turn, influence the duration perception. Research in both adults and children find that knowing that two events are causally related can subjectively compress the temporal delay perception between two events, which is called temporal binding (Blakey et al., 2019; Buehner, 2012). Temporal binding happens in various situations, including when people activate the causal system by themselves, observe the system activated by agents or non-living mechanisms (see Hoerl et al., 2020; J. W. Moore & Obhi, 2012, for review). It reflects the tight relationship between temporal contiguity and causal reasoning in human cognition.

Explanations for this "short-delay preference" consider the normative intuition that the longer the delay, the more likely that alternative causes in between could be responsible for the effect (Lagnado & Speekenbrink, 2010). Another theoretical assumption is that the longer delay between cause and effect, the harder for the information to be sustained in working memory (Buehner et al., 2003; Einhorn & Hogarth, 1986; Ahn et al., 1995).

However, the short-delay preference did not show up in all causal learning studies. In fact, it has only been consistently observed in studies where delays were manipulated within participants.

In contrast, when delays were manipulated between participants, there was no difference in learning performance between those exposed to short delays and those exposed to long delays (Zhang & Rottman, 2021a). Even under the within-subject design, the short-delay preference disappeared when the total duration of observations aligned with the causal delays (e.g. the observation duration of the 6s-delay trials was twice as long as the observation duration of the 3s-delay trials; Lagnado & Speekenbrink, 2010). Neither of the accounts presented above can fully explain this deviation. In Chapter 4, I will propose a third parsimonious account within our Bayesian model that can account for when it is preferable to have short delays in causal attribution.

3.2.3 Predictable delay

People tend to make stronger causal attributions when the temporal delays between a putative cause and effect are less varied across repeating observations (Greville & Buehner, 2010; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Lagnado & Speekenbrink, 2010; Gong et al., 2023). Greville & Buehner (2010, 2016) found that in addition to a main effect of short-delay, people also tend to give higher causal ratings when the delays between cause and effect are drawn from a wider range (e.g. 3-9 s), as opposed to a narrower range (e.g. 4.5-7.5 s) with the same average delay. The unvaried-delay effect persists even when learning duration is increased, suggesting that it is not simply a matter of insufficient learning information (Greville & Buehner, 2010). Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018) asked participants to choose between two causal structures based on multiple clips with consistent activation orders (e.g. $A - B - C$) but variable temporal delays. They found that people preferred the "Chain" structure ($A \rightarrow B \rightarrow C$) when the delay between A and C was relatively variable but the delay between B and C was constant, and preferred the "Fork" structure ($B \leftarrow A \rightarrow C$) when the delay between A and C was constant but the delay between B and C was variable.

To date, there is no clear explanation for the combined effect of short and unvaried delays on causal attribution, as Greville & Buehner (2010) wrote: "Presumably, if a temporal interval is highly predictable, and therefore provides good support for a causal structure model, the extent of delay should not matter". In Chapter 4, I will demonstrate that our Bayesian model can offer a joint explanation for both preferences.

3.3 Order

We have seen how temporal delay plays a role in causal learning. Actually, there is more fundamental thinking of how temporal information help reveal causal relationships: since causes precede their effects. This statement is so intuitive that it is even seen as a definition of causality (Hume, 1740) rather than an assumption that should be tested. Nevertheless, Lagnado & Sloman (2006) design an ingenious task to examine its influence. Participants are asked to imagine a situation

that computer virus can spread through the network and told that the time at which a computer reveals its infection could occur after a variable delay, so later than the time at which the computer became infected. Participants watch multiple clips showing the order of virus appearing in each computer, and then judge the structure of the computer network. Since participants do not know the actual infection time, the presumed solution under CBNs is to summarise contingency information from real-time clips and update beliefs according to the observed contingencies. However, participants' judgments are well-aligned with experienced temporal order rather than the statistical contingency. This suggests that people primarily use temporal orders to make causal inferences, or say, their reliance on order information is so strong that it could not, in this case, be overcome by contingency information (see Chapter 4 for more description of this study).

One interesting question here is whether people think that cause and effect can happen simultaneously. Burns & McCormack (2009) find that 6-7-year-old children strongly favour a common cause structure $B \leftarrow A \rightarrow C$ over a chain structure $A \rightarrow B \rightarrow C$ when B and C happening simultaneously and after A. It shows that children do not favour the answer that contains simultaneous assumptions, even though children are considered to be more open-minded than adults in causal learning (Gopnik et al., 2015; Lucas et al., 2014). Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018) expand this study to include more causal structures and test them on adults. Participants' answers are better predicted by models that assume causes should be not late and also not simultaneous to the occurrence of their effects. Hypotheses failing to satisfy this criterion could be ruled out directly in causal inferences. Although there are situations in the real world that timing differences between cause and effect are undetected by human sensations, such as electronic transmissions, these studies show that people are reluctant to assume the simultaneity when they observe real-time causal events. It could potentially support the assumption that people represent causal events in a continuous timescale where causes and effects cannot happen at exactly the same time.

Similar to temporal binding, there is also a cognitive illusion showing the inverse effect between order and causal direction: causal reordering. People invert the order in which two events happen if they believe the later event to be the cause of the earlier one, and it is better explained by the essential assumption we hold for causality that causes precede their effects than pure attention or memory deficits (Bechlivanidis & Lagnado, 2016). Both temporal binding and causal reordering illustrate that human causal beliefs of temporal information are so strong and robust that they can have top-down influences on the sensation.

3.4 Two other temporal data forms

The research discussed above primarily examines causal relations that manifest as unfolding *events* in continuous time, which will also be the primary focus of this thesis. However, it is important

to acknowledge that there are other forms of data that are relevant to capturing temporal information. I introduce two of them below.

3.4.1 Time-series data

Besides causal learning through events occurring in continuous time, recently several studies have investigated situations that involve *continuous variables* and *continuous timelines*, which are referred to as time-series data (Soo & Rottman, 2018, 2020; Zhang & Rottman, 2021b) produced by continuous dynamic causal systems (Davis et al., 2020; Rehder et al., 2022; Bramley et al., 2019; Btesh et al., 2023). Some of these studies have shown that people can leverage moment-by-moment transitions (i.e. changes in the values of variables between successive observations) to identify the presence and direction of causal relationships (Soo & Rottman, 2018). It has also been shown that people can often identify the causal structure of dynamic systems that involves three continuous variables (Davis et al., 2020; Rehder et al., 2022; Btesh et al., 2023) if they can freely intervene on and control each variable in real time. Participants in these tasks appeared to follow an intervention strategy of creating occasional dramatic and rapid changes in variables and monitoring the behavior of other variables shortly afterward. They were able to learn better when the variables changed rapidly and affected one another rigidly rather than slowly or gradually (Rehder et al., 2022; Gong & Bramley, 2022), which is aligned with the short-delay preference discussed above when people infer from events.

Granger causality (Granger, 1969) is an established statistical technique designed to identify potential causality in time-series data, with a mechanism to accommodate causal lag. To assess if one variable "Granger causes" another, one searches across a range of fixed lags deemed to be mechanistically plausible, e.g. $X_{t-1} \dots X_{t-m}$, and tests whether inclusion of any of these terms statistically improve prediction of Y_t over and above its own lagged autocorrelation (modeled by including $Y_{t-1} \dots Y_{t-m}$ as a covariate). If a statistical relation is found for one or more of these lags, the causal influence is deemed to be supported. As such, Granger causality does not inherently privilege longer or shorter lags. This lag-indifference may be appropriate for minimizing bias when modeling domains that are poorly understood but may not reflect human expectations. Indeed, people more reliably identify a relationship when its causal lag is short than long (Gong & Bramley, 2022), which is consistent with the context of event-based temporal causal learning I introduced above.

3.4.2 Spatiotemporal data

Researchers examine spatiotemporal evidence when asking individuals to make causal inferences for 2D physical scenes (Ullman et al., 2018; Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018; Gerstenberg et al., 2021). Due to the brevity of the clip, it is crucial to leverage the mechanistic details to discover the underlying structure. This necessitates that learners consider

the specific movements of objects that are causally related to one another. Ullman et al. (2018) use a hierarchical Bayesian framework to demonstrate this normative process. Reasoners must hold both higher-level, abstract principles and detailed mechanistic knowledge to effectively reason about spatiotemporal evidence. The higher-level causal conclusions are finally draw from the investigation of individual evidential details.

Nevertheless, the complexity of spatiotemporal dynamics can introduce the problem of intractability (Griffiths, 2020). In reality, it is difficult for people to follow normative frameworks, and hence they often rely on approximations by relying on simulations to extract heuristic cues and examine cues in the evidence (Ullman et al., 2018; Bass et al., 2021). As a result, inferring hidden causes in physical scenes can also be challenging. C. D. Carroll & Kemp (2015) presented participants with only the effect and asked them to infer the hidden causes that influenced its motion. They found that participants often failed to generate the location of hidden causes, even though they could endorse it when it was presented to them. This reveals the difficulty of inferring hidden causes when mechanistic information is complex and the hypothesis space is large. As such, even with rich spatiotemporal data, it may be challenging for people to capture all the details of the information (Rehder et al., 2022; Ludwin-Peery et al., 2020, 2021). I will show in later chapters that the similar computation issue exist in event-based temporal causal learning tasks and it would be taken into account when I develop the models to describe human performance.

Chapter 4

Rational causal induction from time

In the previous chapters, we have seen the history of causal theories and the evidence of people's sensitivity to temporal information in causal reasoning. However, so far, we do not have a quantitative theory about how temporal information should be employed systematically for the purpose of learning causal structures. In this Chapter, I develop a rational framework that incorporates the role of time in guiding causal learning. I define a formal framework for expressing continuous-time causal theories, particularly attending to the role of time in the construction of these theories. I work within the Bayesian rational analysis tradition (Marr, 1982; Anderson, 1990), as this has proven successful in developing theories of atemporal causal induction (Griffiths & Tenenbaum, 2005; Rottman & Hastie, 2014). However, I depart from past analyses of causal inference by linking causal influence with dependence between events in continuous time (i.e. *contiguity*) rather than co-incidence of variable *states* across independent trials (i.e. *contingency*). I will show how this approach anticipates the ceteris paribus human preference for causal explanations that posit shorter, more reliable and more predictable causal influences. Furthermore, I will show this account provides a unified explanation for human judgments across seven experimental datasets from the causal learning literature.

Content from this chapter is based on a project collaborated with M Pacer (co-leader), Thomas L. Griffiths, and Neil R. Bramley.

4.1 The variety of temporal causal learning scenarios

Temporal evidence utilized for learning causal structures can manifest in various forms. For instance, a rain shower can be perceived as a singular event or alternatively as a sequence of numerous raindrops, which are essentially multiple events occurring one after the other. In terms of causal relationships, we might think of an effect as a specific event, whereas in other cases, we might be focused on the change in the number or rate of density of events of a particular class or type.

We further illustrate this idea using a simple example in Figure 4.1. We suppose a fictional substance called 5-HTP is used to treat insomnia. Consuming a 5-HTP capsule can cause a person to sleep, representing a one-cause–one-effect scenario. Here, temporal information is embedded within the *delay* between the causative event of pill consumption and its effect event of falling asleep.[1] The causal delay can vary across different mechanisms and hence follow distributions with different shapes (see Figure 4.1a), analogous to our anticipation of certain medications (e.g., Adrenaline) taking effect rapidly and precisely, while others (e.g., painkillers) exhibit a delayed onset with some degree of variability. However, we might also model the same effect more granularly in terms of the pill's production of Melatonin particles over time (Figure 4.1b). One-cause–many-effect scenarios are prevalent in epidemiology. For example, a single water pollution event might cause many individuals to fall ill at different points in time (Griffiths & Tenenbaum, 2007). In this case, instead of focusing on the relationships between the cause and individual effect events, it may be more practical to think at a macro-level about how the cause affects the *rate* of particle production, or illness, how and whether this rate departs from (and later returns to) its base rate. This requires reasoning about the functional form of the event's causal influence in time, including a potential incubation period, peak, and a decay process (see Figure 4.1b).

From a cognizer's point of view, in either situation above, the delay between a specific cause and its putative effect is liable to be *variable* and uncertain. This is an inevitable feature of any model that abstracts away some of the detail, leaving unmodeled noise and complexity in the generative or measurement processes. Therefore, a rational model of time-based causal induction needs to capture how abstract subjective probability distributions encode causal-model-based expectations about inter-event delays and rates, and how these distributions can be shaped and sharpened with evidence.

The process of collecting temporal evidence can also vary depending on the context. All evidence could be collected from a single causal system, where all events occur within a single timeline, as depicted in Figure 4.2a. For instance, in Lagnado & Speekenbrink (2010), participants observed a geological system for several minutes, tracking the occurrence of wave events (potential causes) and earthquake events (the effect) unfolding over time. Alternatively, evidence can also be gathered from multiple independent samples. For example, in the research conducted by Greville & Buehner (2007), participants observed the timing of the death of each bacteria culture sample (the effect) after receiving a particular treatment (the cause). As shown in Figure 4.2b, instead of having multiple cause and effect events within a single timeline, there is one cause event and one effect event in each timeline, with multiple timelines obtained from different independent samples. In the rest of the paper, we refer to the former situation as "continuous evidence", while the latter is termed "episodic evidence".

[1] We here have not delved into real scenarios that consider base rate events, such as individuals naturally falling asleep. We will explore situations that encompass multiple causes and base rate effects in subsequent sections.

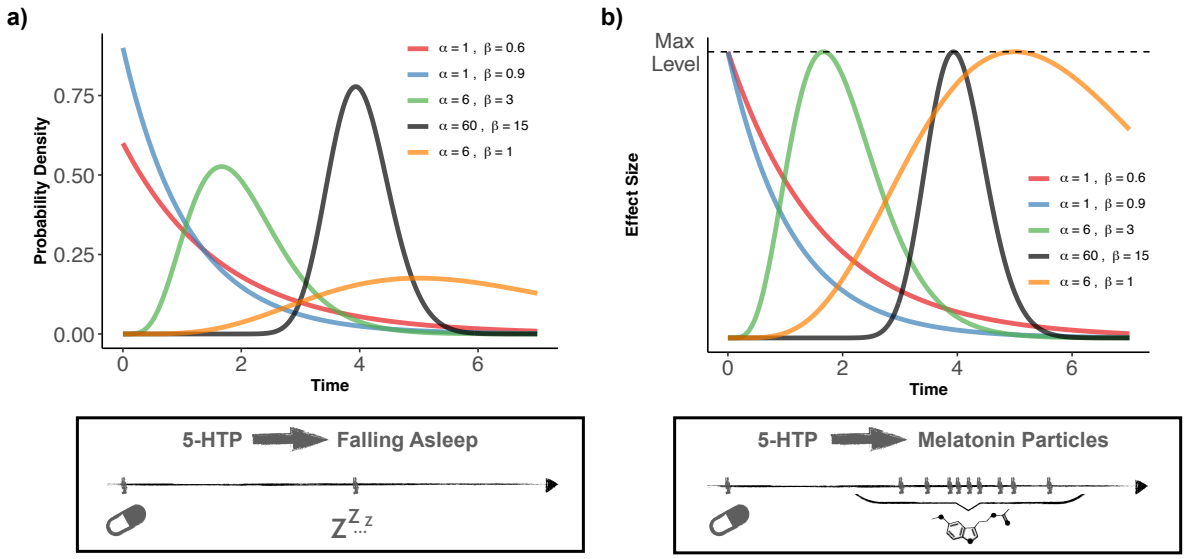
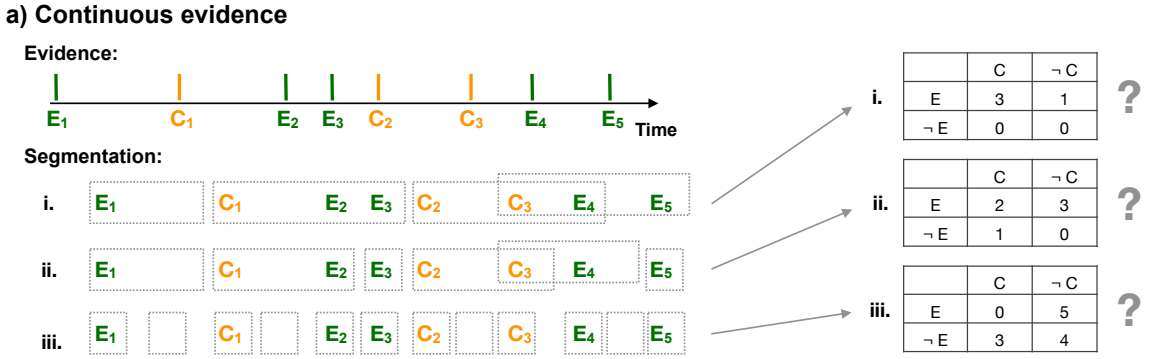


Figure 4.1: Examples of two types of function that could be used to model cause-effect delays and causal influences, respectively. Illustrative example relates a drug "5-HTP" and sleep. a) Gamma probability density function capturing delay between drug and sleep and b) scaled gamma density function capturing the rate of melatonin production after drug is administered.

We can see that both continuous evidence and episodic evidence pose challenges when it comes to encoding them in contingency tables and applying atemporal causal calculations (Cheng, 1997; Griffiths & Tenenbaum, 2005; Perales & Shanks, 2007). Continuous evidence presents the difficulty of making arbitrary decisions when combining events, resulting in a multitude of contingency tables that could yield inconsistent causal judgments (see Figure 4.2a). On the other hand, episodic evidence can be quantified by merely observing whether the cause and effect occur within a timeline, but this fails to capture the significance of the temporal delay between cause and effect.

In addition to the distinction between continuous and episodic evidence, differences in temporal causal learning tasks can also arise in other domains, such as whether the effect variables are specified or not. In cases where the effect variables were specified, participants were asked to identify the causes of a particular effect variable. In other cases, participants were tasked with determining the existence of a connection between two variables and the causal direction of that connection. We will model seven human datasets that were categorized into four groups based on the nature of the evidence (continuous or episodic) and whether the effect variables were specified, as shown in Table 4.1. Meanwhile, these datasets also have variations in other characteristics, including: (1) base rate: whether the effect could occur without any endogenous causes; (2) prevention: whether preventative causal relationships were considered; (3) cycle: whether cyclic relationships were taken into account; and (4) delay expectation: whether participants were informed or trained about causal delays prior to the task.



b) Episodic evidence (Greville & Buehner, 2007)

Experimental Group

	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5
1	C				E	
2	C		E			
3	C	E				
4	C			E		
5	C					E

Control Group

	Day 0	Day 1	Day 2	Day 3	Day 4	Day 5
6						E
7		E				
8						
9				E		
10						

Figure 4.2: a) Examples of the possible arbitrary decisions when segmenting continuous time evidence into contingency evidence, along with the corresponding contingency tables. b) Examples of episodic evidence adapted from Greville & Buehner (2007). In the experiment, participants assessed the impact of a treatment (C) on the survival of bacterial cultures, considering culture death as the outcome (E).

We will demonstrate that our rational framework can deal with the aforementioned variations in temporal causal learning tasks. Meanwhile, given that those tasks have been tested empirically, we will demonstrate the degree of sensitivity people exhibit towards the rational framework.

4.2 Formal framework

Beginning this formalization effort, Griffiths & Tenenbaum (2009) highlight three critical components of a rational causal induction account: (1) An ontology that outlines the entities under investigation and their properties, (2) a set of plausible relations that suggest how entities may be connected, and (3) the functional form that determines how causes influence their effects under each type of relation. Working with causal Bayesian networks (Pearl, 2000; Rottman & Hastie, 2014), we can interpret this process as requiring decisions about what constitute the nodes (i.e. entities) within the causal structure of interest, the hypothesis space that includes combinations of directed edges (i.e. connections) between nodes, and the specification of functional forms used to calculate the likelihood to evaluate the proposed structures. Despite differences in the data they operate over, temporal and atemporal causal induction share a similar process of identifying a hypothesis space of causal structures. The learner updates their prior belief over structures s in

Table 4.1: Dataset features.

Name	Reference	Base Rate	Prevention	Cycle	Delay Prior
Continuous, effect specified:					
Earthquake	Lagnado & Speekenbrink (2010)	✔	✗	✗	✗
Device: Prevention	Gong & Bramley (2023a)	✔	✔	✗	✔
Continuous, effect unspecified:					
Device: Active Learning	Gong et al. (2023)	✔	✗	✔	✔
Episodic, effect specified:					
Bacteria	Greville & Buehner (2007)	✔	✔	✗	✗
Future Bacteria	Gong & Bramley (2023b)	✔	✔	✗	✗
Episodic, effect unspecified:					
Computer Virus	Lagnado & Sloman (2006)	✗	✗	✔	✗
Device: Chain or Fork	Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018)	✗	✗	✗	✗

Note: The human data were from Experiment 2 in Lagnado & Speekenbrink (2010), Experiment 1 in Gong & Bramley (2023a), Experiment 1 in Gong et al. (2023), Experiment 1 in Greville & Buehner (2007), Experiment 1 and 2 in Gong & Bramley (2023b), Experiment 1 in Lagnado & Sloman (2006), and Experiment 3 and 4 in Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018).

the hypothesis space $P(s)$ with a likelihood function $P(\mathbf{d}|s;\mathbf{w})$ to get the posterior distribution $P(s|\mathbf{d};\mathbf{w})$, given data $\mathbf{d}$ and a set of parameters $\mathbf{w}$ [2]:

$$P(s|\mathbf{d};\mathbf{w}) \propto P(\mathbf{d}|s;\mathbf{w}) \cdot P(s) \tag{4.1}$$

Here, we aim to address the critical question of determining the appropriate functional forms to calculate the likelihood $P(\mathbf{d}|s,\mathbf{w})$.

We here employ Gamma distributions as the probabilistic distributions to model causal delays (Figure 4.1a). Gamma(α,β) defines a density over $(0,+\infty)$ with two parameters, shape α and rate β, which control the expectation and central tendency of the delay (mean $\mu=\alpha/\beta$ and variance $\sigma^2=\alpha/\beta^2$). The use of Gamma distributions offers several advantages over other probabilistic distributions (e.g. normal or log-normal distributions), including (1) its range naturally aligns with the definition that causal delays should be a positive real number; (2) the shape and rate parameters can capture a wide range of possibilities. Box 4.1 at the end of this chapter also summarizes several desirable mathematical properties.

How can gamma distributions be used to model different causal mechanism? In the one-cause–one-effect cases, the duration of the causal delay may be uncertain, and the delay could vary widely among repeated observations. This delay uncertainty P_d can be represented using a

[2]We here focus on the problem of structure selection rather than parameter estimation (Griffiths & Tenenbaum, 2005). That is, we theoretically assume that for each structure the parameters are marginalized from the entire range of possibilities if they are unknown.

Gamma probabilistic density function (as shown in Figure 4.1a) :

$$P_d(t|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t} \tag{4.2}$$

Exponential distributions are special cases of Gamma distributions when the shape parameter $\alpha = 1$ (see Figure 4.1a). In this case, the delay between cause and effect events is "memoryless" property, meaning that the probability of waiting another unit of time Δt to see the effect is constant, and therefore independent of how long one has already been waiting (see Box 4.2 at the end of this chapter for the proof of the memoryless property). This property is useful for modeling spontaneous effects (or say effect events caused by hidden causes) since they are often unpredictable, and there is no privileged time at which to start measuring how long should expect to wait, their occurrence is as likely at any moment as at any other.[3]

For the one-cause–many-effect cases, we can construct the function to capture causal influence dynamics I by scaling the Gamma density function via dividing by its mode, i.e. the density at $(\alpha - 1)/\beta$:

$$I(t|\alpha,\beta) = \frac{P_d(t|\alpha,\beta)}{P_d(\frac{\alpha-1}{\beta}|\alpha,\beta)} \tag{4.3}$$

After scaling, the predicted value ranges from 0 to 1, where 1 means the causal influence reaches its maximum level (see Figure 4.1b). We adopt this form here as a mechanistically agnostic default for simplicity but recognize that, in principle the influence of a cause on the rate of an effect could have any functional form. For instance, the causal influence size may remain at its peak level for an extended period before decaying. We will discuss this situation in one of the datasets later on (Gong & Bramley, 2023a). However, we believe utilizing the function above is an effective approach to capture many cases in previous experiments (Gong et al., 2023; Greville & Buehner, 2007; Lagnado & Speekenbrink, 2010; Lagnado & Sloman, 2006), and in other contexts the functional form could be derivable from mechanism knowledge.

4.2.1 Event–based and rate–based schemes

The two examples presented above suggest that there are at least two closely related approaches for reasoning about temporal evidence. One approach involves inferring the causal relationship by considering (1) the delay between the proposed cause-effect pair, while the other involves examining (2) how the effect rate changes after the cause occurs. The former, *event-based scheme* estimates the likelihood using the causal delay function $P_d(t|\alpha,\beta)$, while the latter, *rate-based scheme* estimates the likelihood using the causal influence function $I(t|\alpha,\beta)$.

[3]Note that another special case of the gamma distribution occurs when the shape parameter $\alpha < 1$, resulting in the expectation that the effect will happen very soon or very late, a little like an inverted Gaussian with all its mass at the tails, and none at the mean. While this case may be appropriate for some scenarios, we do not include them here.

In contrast to the 5-HTP examples, real-life continuous-time evidence can be much more complicated: Events can occur at any time point, and different causes can interweave over time to influence the outcome. For instance, imagine that you are taking a pill for medical purposes but frequently experience stomach discomfort afterward. You might wonder whether this discomfort is a side effect of the pill. We illustrate this evidence in Figure 4.3a. The timing of the pill ingestion could be arbitrary, and the effect could occur multiple times during the observation period. If the causal relationship exists and the causal delay is long, the stomach discomfort produced by one pill could happen after the ingestion of another pill. Therefore, it is impossible to divide the evidence into independent trials. In this example we focus on two hypothetical structures S_0 and S_1 in Figure 4.3b. In S_0, only the base rate B causes the discomfort, while in S_1, both the base rate B and the pill taking C cause the discomfort. In other situations, if the learner suspects that other factors, such as diets, may also contribute to the stomach discomfort, they may need to include additional diet events in the timeline, which can further complicate the evidence. It is necessary to preprocess the data to calculate the likelihood under the causal delay or influence function.

Event-based scheme

The event-based scheme uses the concept of token-level "actual causation" to map each event to its possible causes (Halpern, 2016), identifying which of several candidate events actually caused the observed outcome (Stephan et al., 2020; Gerstenberg et al., 2021). While we may have knowledge and expectations about the delay between a cause and its effect (i.e. the mean and variance), to utilize these directly we have to also commit to a particular causal story about which cause event actually produced which effect event in order to apply those expectations. Under this scheme one can consider various possible causal pathways that could produce the observed events, depending on the underlying causal mechanisms (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Gong & Bramley, 2020; Valentin et al., 2020). For example, in a causal structure S that includes an endogenous cause C, an hidden background cause B, and an effect E, each effect event could be caused by either C or B, resulting in a total of 2^k possible pathways in the set $(\mathbb{Z}_s)$, where k is the number of effects. The event-based scheme allows for specific mechanistic constraints to be integrated into pathway construction. For instance, if we observe a sequence of events, such as $\{C_1, E_1, E_2\}$, and also believe that this is the kind of system within which one C event can only cause one E event, we can rule out the pathway that assumes both E_1 and E_2 were caused by C_1.

Given that conditional on a structural hypothesis, the potential actual causal pathways are mutually exclusive and exhaustive, it follows that the overall likelihood of each structure hypothesis is the sum of the individual likelihood of these pathways:

$$P(\mathbf{d}|s;\mathbf{w}) = \sum_{\mathbf{z}\in\mathbb{Z}_s} P(\mathbf{z}|s;\mathbf{w}) \tag{4.4}$$

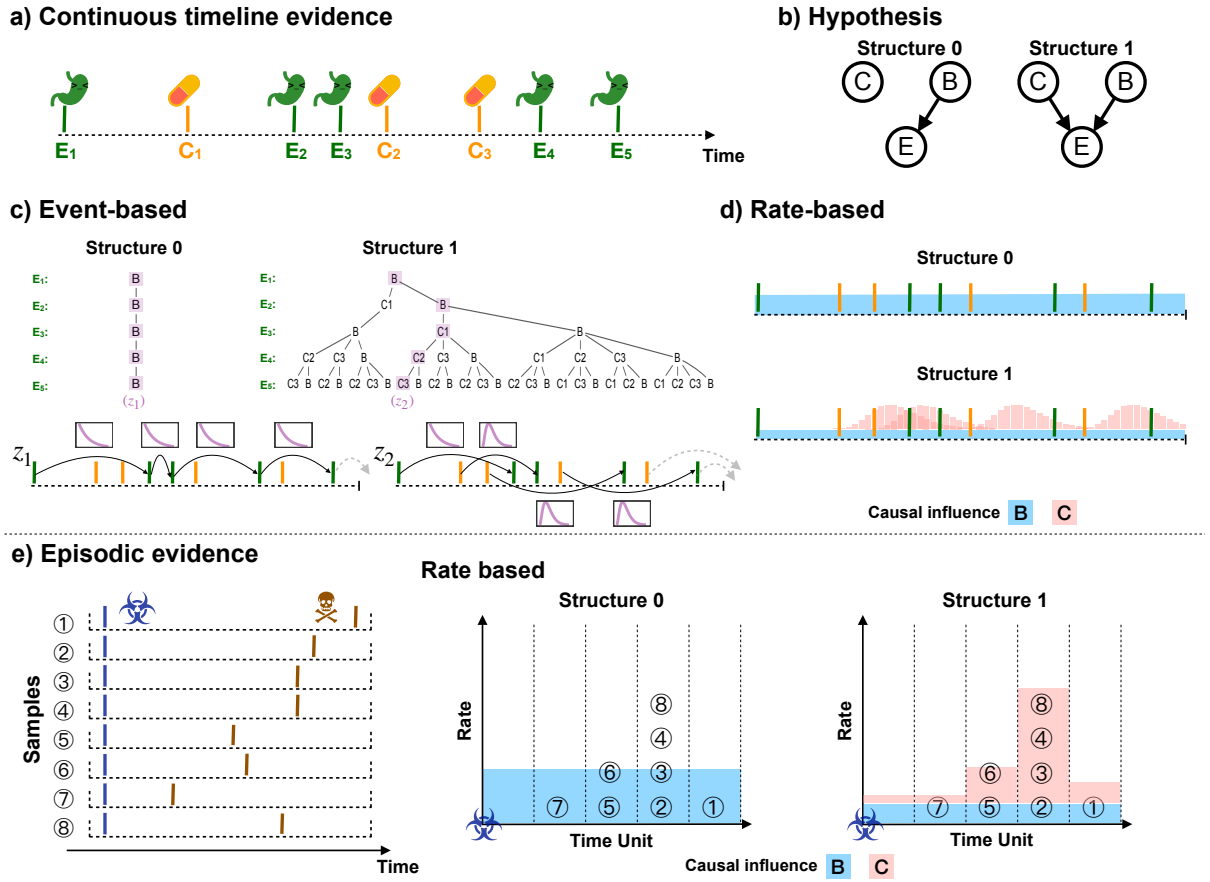


Figure 4.3: Causal inferences based on continuous-time causal evidence. a) Evidence as Events of stomach discomfort and pill taking unfolded in the timeline. b) There are two causal structures in the hypothesis space. c) The event-based scheme lays out all possible pathways (branches) that explain all effects under each hypothetical structure. d) The rate-based scheme model in what way the rate of effects are expected to change under each hypothetical structure. e) Episodic type of evidence where the cause and effect only happen once in each individual observation. Cases illustrated the situation in (Greville & Buehner, 2007) where the effect events across samples are assumed to follow exponential delays if the evaluated cause does not work. Under this situation, the evidence can be collapsed under the rate-based scheme.

To determine the likelihood of each pathway $P(\mathbf{z}|s;\mathbf{w})$, we analyze both existing effect events e and hidden effect events h. For each existing effect e, we evaluate the probability that (1) it was caused by the presumed generative cause event g as well as that (2) it was not prevented by a set of presumed preventative cause events $\mathbf{p}$. Hidden effect events occur when we cannot identify the effect event of a generative cause. This could be due to (1) the generative cause failing to produce the effect, (2) the effect being prevented, or (3) the effect not having occurred yet:

$$P(\mathbf{z}|s;\mathbf{w}) = \prod_{g\to e\in\mathbf{z}} \underbrace{w_g \cdot P_d(t_e - t_g|\alpha,\beta)\cdot(1-P_p(e))\cdot}_{\text{Observed effects should be generated and not prevented}} \prod_{g\to h\in\mathbf{z}} \underbrace{(1-w_g)+w_g\cdot P_d(t_h > t_{end}|\alpha,\beta)+w_g\cdot P_p(h)}_{\text{Unobserved effects should be not existing, not unrevealed yet, or prevented}} \tag{4.5}$$

The event-based scheme provides flexibility in dealing with preventative causation $P_p(e)$ (the probability that e should have been prevented) based on different rules. For instance, the preventative cause can block effects for a specific time window or block the nearest effect. It can also block all effects equally or selectively block effects from a particular cause (C. D. Carroll & Cheng, 2009; Gerstenberg & Stephan, 2021). We will demonstrate these different possibilities using a dataset (Gong & Bramley, 2023a) later on.

In Figure 4.3c, the event-based scheme generates pathways for explaining stomach discomfort under different structure hypotheses. For S_0, all effect events are attributed to the base rate. For S_1, any effect event can be attributed to the base rate or any cause events that occurred previously. We put the constraint that each cause event causes only one effect event. We illustrate two pathways $\mathbf{z}_1$ and $\mathbf{z}_2$ as examples. The base rate event is represented as being caused by the previous base rate effect (making its process independent from the evaluated cause). We usually model the base rate delay using memoryless exponential distributions if it is unpredictable. The unobserved events, represented by dashed arrows, occur when we cannot find the corresponding effects of a cause in the timeline. This can happen when the evaluated cause fails to work or when its effect has not yet occurred. For the base rate cause, this can occur when the next base rate has not yet happened. This is due to that we do not need to assume a failure rate of the base rate when the timing of it is already unpredictable.

Rate-based scheme

The rate-based scheme does not account for a one-to-one mapping between individual cause and effect events. In contrast, it is concerned with the change in the rate of effects following the occurrence of a cause event. When the cause is generative, the rate of effects typically increases, whereas preventative causes tend to decrease the rate of effects. As such, the scheme shifts the focus from delay to rate, making the Gamma probability distribution unsuitable for modeling. Instead, the Poisson process can be used to capture the probability of observing a particular rate $P_r(k)$ (the number of events at a time unit) given a rate assumption:

$$P_r(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4.6}$$

We assume a base rate of an effect as a constant λ_0 given that no information is available to suggest how it could change across time. For any generative cause, the causal influence after the cause occurrence can be modeled as an incubation-decay process as shown in Figure 4.1b, captured by the influence function $I(t|\alpha, \beta)$. The generative rate can be represented by another function of time $f(\lambda_1, t) = \lambda_1 \cdot I(t|\alpha, \beta)$, where λ_1 denotes the maximum level of causal inference. Preventative causes, on the other hand, are presumed to decrease the effect rate by a proportion ranging from 0 to a maximum level of ξ ($0 < \xi < 1$). The preventative influence can also follow an incubation-decay process and be represented by a function of time $f(\xi, t) = \xi \cdot I'(t|\alpha, \beta)$.

Poisson processes have a desirable property known as "superposition", where the union of two independent Poisson processes with rates λ and λ' is still a Poisson process with rate $\lambda + \lambda'$. Conversely, preventative causation can be viewed as "thinning" processes that selectively filter out some effect events with a probability of ξ'. Combining multiple causes with a base rate of λ_0, the expected effect rate $f(\lambda, t)$ at the time unit t can be represented using the noisy-OR and noisy-AND-NOT principles by accounting for superposition and thinning as follows:

$$f(\lambda, t) = (\lambda_0 + \sum_{i \in \mathbf{g}} f(\lambda_i, t)) \prod_{j \in \mathbf{p}} (1 - f(\xi_j, t)) \tag{4.7}$$

The superposition and thinning properties not only give us a simple answer to the combination of a base rate and a (constant) causal influence, but also how a non constant causal influence implies a fluctuating rate. When the rate of the Poisson process can changes across time, it is called the non-homogeneous Poisson process. The likelihood depends on how the observed rates at each time bin are aligned with the expected rates:

$$P(\mathbf{d}|s; \mathbf{w}) = \prod_t P_r(d_t | f(\lambda, t)) \tag{4.8}$$

Figure 4.3d illustrates how the rate-based scheme generates expected rate changes to explain stomach discomfort. In S_0, the model assigns a constant base rate to account for the number of effect events per unit of time. In S_1, the model incorporates the assumption that the effect rate dynamically changes following the occurrence of a cause event.

Summary and comparisons of two schemes

Event-based and rate-based schemes differ in the granularity of their focus, leaning in one case more *token-level* and in the other case more towards *type-level* causal reasoning. Type-level thinking involves considering the causal structure relating event types (i.e. which type of class causal event influences which type of effect events). Token-level thinking involves considering the actual causal relationships between particular events (i.e. which particular cause event truly caused which particular effect event). We present this dual view because, depending on the inference setting, one or other mode is more appropriate. However, we will show how in most cases, both modes make similar predictions and perform similarly well in capturing human judgments, and highlight cases where computational and representational costs differ dramatically making one algorithm more suitable than the other.

Each scheme has its own set of strengths and limitations when dealing with specific scenarios. For instance, the event-based scheme is capable of incorporating periodic base rate knowledge by designating certain events as base rate events and employing predictable delays (gamma distributions) to model the intervals between them. Additionally, it can effectively handle mechanistic

knowledge, such as cases where a single cause produces (or prevents) a single effect. In contrast, the rate-based scheme does not distinguish the source of each effect event, rendering it incapable of tracking base rate events or determining when to stop expecting a specific effect event from a given cause.

However, the requirement for actual causation is not always preferable. One obvious reason is that when a single cause can lead to multiple indistinguishable events, attempting to determine which specific events were caused by which factors can significantly amplify computational complexity without providing substantial benefits. From a computational cost perspective, the event-based model proves to be highly demanding due to the necessity of considering all possible combinations of cause and effect events. This results in a greater than polynomial increase in computational cost as the number of events grows. Conversely, the computational cost of the rate-based model exhibits *linear* growth with the observation duration, once the time unit granularity is chosen. Additional reasons become apparent when examining a case study (Greville & Buehner, 2007) later on.

4.3 Human generic delay principles: short, predictable, and expected

Humans show preferences for short, predictable, or expected delays as a suggestion for stronger causal attributions (see Chapter 3). We here demonstrate that these preferences could be explained from a rational Bayesian perspective. We first provide intuitive explanations for these phenomena and then provide simulation results to confirm. Each generic principle found in humans can be manifested in two types of tasks: structure induction and diagnostic causal reasoning. For example, the principle of short delay can refer to learners' (1) higher causal ratings for an evaluated cause in trials where the delays are short, as compared to when they are long, or (2) preference for Cause X over Cause Y as an explanation for the effect in a single learning trial. We focus primarily on the first scenario but demonstrate how the same principles apply readily to the second task.

Even in the absence of delay expectations, a Bayesian preference for short delays can be understood from two perspectives. The first perspective arises from the fact that, given causal delays can range from zero to infinity (with a limit on the lower side but not the upper side), delay distributions tend to be more or less right-skewed. As such, it can often assign more probability to a smaller number than a larger number when they have the same distance from the central tendency (the mean value) of the distribution. As shown in Figure 4.4, under distributions with an expected delay of 4 seconds, a delay of 2 seconds often receive more likelihood than the delay of 6 seconds, unless the variance become small. In other words, short delays often exhibit higher probability density than long delays, and these advantages can accumulate across different prior beliefs sampled from the prior distribution. The second reason stems from the absence of delay

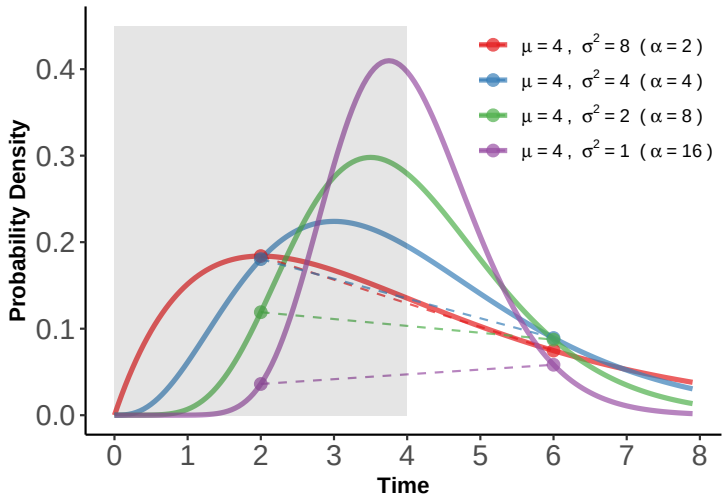


Figure 4.4: Illustration of how Gamma distributions favor short delays (a higher density for the 2-second delay than the 6-second delay) when the uncertainty (variance) is high due to the right-skewed property.

expectations, resulting in a bigger range of larger variances for long delay expectations compared to short delay expectations, making long delays harder to predict accurately.

The variability principle can also be explained by the *likelihood* calculation. When delays exhibit greater variability, it becomes less probable for a cause to receive a high likelihood under any specific gamma distribution, leading to a lower posterior probability compared to situations with less variable delays.

The expected-delay principle can be understood as the influence of mechanistic knowledge or prior experience on people's *prior* distribution regarding the causal delay. For instance, if individuals strongly believe that a genuine switch should take approximately 4 seconds to turn on a device, a switch that takes 2 or 6 seconds would have a lower prior probability and consequently a lower posterior probability compared to a switch that takes 4 seconds. That is, the device activation after 2 or 6 seconds would be more likely to be accounted by the base rate rather than the switch (Buehner & McGregor, 2006).

4.3.1 Simulation

We here simulate data via the following procedure. In order to demonstrate how our model can handle data that is not exclusively generated from gamma distributions (just like humans, Greville & Buehner, 2010), we utilize uniform distributions, denoted as $U(l, u)$, with a lower bound l and an upper bound u:

1. Assuming evidence lasts for 300 time units. Cause events C occur k_c times sampled from $U(0, 300)$.

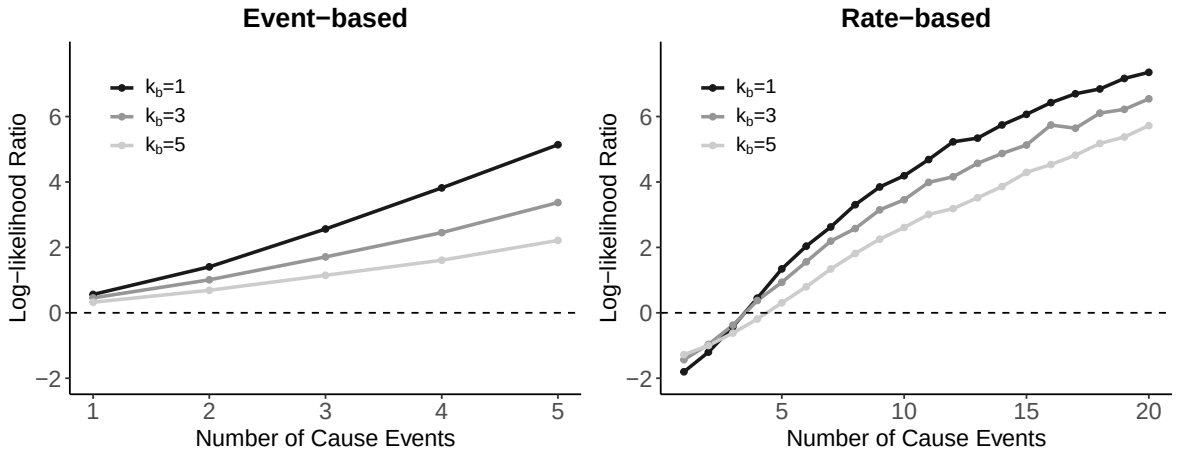


Figure 4.5: How the log-likelihood ratio changes with the amount of cause events.

2. Each cause event has a probability w_c of producing one effect event E with the delay between them sampled from $U(m_u - i_u, m_u + i_u)$.

3. Another k_b base rate effect events are sampled from $U(0, 300)$.

The learner is asked to judge how possible that there is a generative link between C and E, i.e. judging S_0 and S_1 in Figure 4.3b. The evidence that data $\mathbf{d}$ provide in favor of S_1 over S_0 can be calculated as log-likelihood ratio, assuming both structures have equal prior probabilities (Griffiths & Tenenbaum, 2005):

$$log \frac{P(\mathbf{d}|S_1; \mathbf{w})}{P(\mathbf{d}|S_0; \mathbf{w})} \tag{4.9}$$

We assume that all parameters used for simulating data are unknown to the model and hence the model need to marginalize over the parameter set $\mathbf{w}$ to estimate the likelihood of a causal structure. We use Monte Carlo simulations with sample size $m = 10,000$ to approximate the Bayesian inference. For event-based scheme, it assumes the cause succeeds to produce the effect with a probability of $w_c \sim U(0, 1)$ and the causal delay between C and its effect E follow a gamma distribution with mean $\mu \sim U(0, 300)$ and variance $\sigma^2 \sim U(0, \mu^2)$. This ensures that the prior around the mean is weak (0-300 is a very large range given that each observation only lasts for 300 time units) and that the shape parameter will be larger than 1 ($\alpha = \frac{\mu}{\sigma^2}$). [4] Similarly, the delay between two base rate events follows an exponential distribution with mean $\mu_b \sim U(0, 300)$. For the rate-based model, we specify the base rate $\lambda_0 \sim 1/U(0, 300)$, the max causal influence $\lambda_1 \sim U(0, 1)$. The causal influence changes dynamically given a gamma distribution with mean $\mu \sim U(0, 300)$ and variance $\sigma^2 \sim U(0, \mu^2)$. We summarize the meaning of symbols used in Table 4.2.

[4]One alternative approach is to sample the mean (μ) and shape (α) parameters from very flat exponential distribution (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018). This returns similar results.

Table 4.2: Symbols used and their meanings under three contexts in this chapter.

	Simulation process	event-based scheme	rate-based scheme
m_u	Mean of causal delays.	–	–
i_u	Half interval of causal delays.	–	–
k_b	Number of base rate events.	–	–
k_c	Number of cause events.	–	–
w_c	Cause's success probability.	Cause's success probability.	–
μ	–	Mean of causal delays.	Mean of causal influence function.
σ^2	–	Variance of causal delays.	Variance of causal influence function.
μ_b	–	Mean of base rate delays.	–
σ_b^2	–	Variance of base rate delays.	–
λ^0	–	–	Effect's base rate.
λ^1	–	–	Max generative causal influence.
ξ	–	–	Max preventative causal influence.

Note: σ_b^2 only appeared when modeling Gong & Bramley (2023a) which included periodic base rates. In other cases, the base rate delay was modeled using exponential distributions, which only included one parameter.

Data points

We first examine the number of data points required for the model to favor S_1 over S_0. We here use $w_c = 1$, $m_u = 15$, $i_u = 5$, and consider different values for k_b ($k_b = \{1, 3, 5\}$). For the rate-based model, we search for values of k_c ranging from 1 to 20 (with a step of 1), while for the event-based model, we limit the search to values from 1 to 5 due to computational constraints. Results are shown in Figure 4.5. The event-based model starts favoring S_1 even with just one cause event. The rate-based model starts favoring S_1 with five cause events, indicating a higher requirement for data points to support S_1 due to the more relaxed constraints of the model. In both cases, the log-likelihood ratio increases as the number of cause events increases. Both models perform better when the base rate is low, which is aligned with atemporal learning setting that learners can better learn a generative relationship when the base rate is low (Cheng, 1997; Griffiths & Tenenbaum, 2005; Wu & Cheng, 1999).

Delay duration and variance

To illustrate the short-delay principle found in humans, we simulate stimuli arranged in a grid with $w_c = \{0.6, 0.8, 1\}$, $m_u = \{10, 15, 20, 25, 30\}$, $i_u = 1$, and $k_b = 3$. Based on the simulation results above, we use $k_c = 5$ for the event-based model and $k_c = 10$ for the rate-based model. Figure 4.6a demonstrates that the event-based model's preference for S_1 over S_0 diminishes as the duration of the delay increases. This observation supports the notion that causal attribution is stronger when the delay is shorter. Additionally, the log-likelihood of S_1 itself decreases as the delay duration increases. This indicates that when faced with multiple potential cause candidates, the learner tends to attribute the effect to the cause with the shortest delay. Similar patterns are replicated in the rate-based scheme (Figure 4.6b).

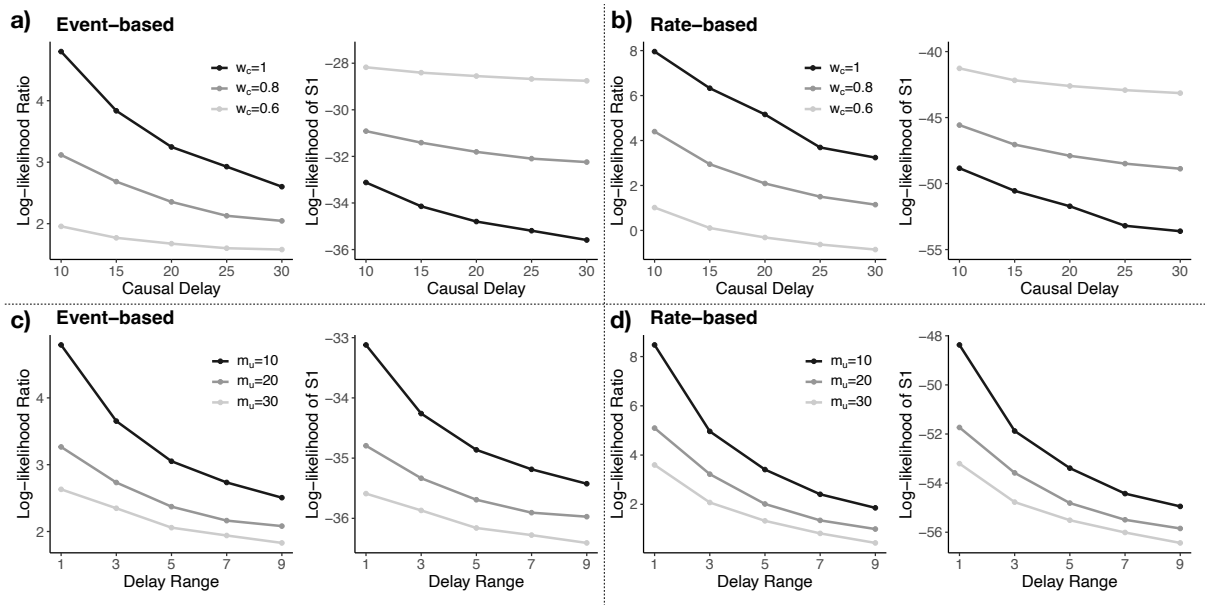


Figure 4.6: How the log-likelihood ratio changes with the causal delay duration and variance.

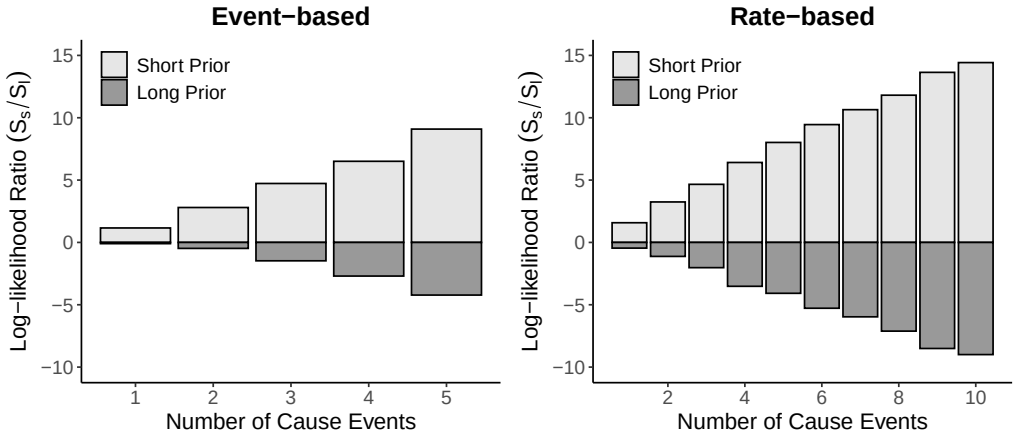


Figure 4.7: How the log-likelihood ratio changes under different delay prior.

To investigate the predictable-delay principle, we simulate stimuli arranged in a grid with $m_u = \{10, 20, 30\}$, $i_u = \{1, 3, 5, 7, 9\}$, $w_c = 1$, $k_b = 3$. Similarly, we use $k_c = 5$ in the event based model and $k_c = 10$ in the rate based model. As shown in Figure 4.6c, the event-based model's preference for S_1 over S_0 diminishes as the range of delays increases. It explains why causal attribution is stronger when the delays are unvaried. Additionally, the log-likelihood of S_1 decreases as the delay range expands, which suggests that when faced with multiple potential cause candidates, the learner tends to attribute the effect to the cause with the most consistent or unvaried delays. Similar results are observed in the rate-based scheme (Figure 4.6d).

Delay expectation

To investigate the influence of prior beliefs on causal judgments, we introduced two different delay prior conditions instead of using the above uniform delay priors. For the "short prior", we set μ to be sampled from a Gamma distribution with a mean of 10 and a standard deviation of 1, resulting in an assumed delay expectation of 10±1. Conversely, for the "long prior", we assume μ is sampled from a Gamma distribution with a mean of 20 and a standard deviation of 1, representing an assumed delay expectation of around 20±1. Other model parameterizations remains the same as previous. For stimuli, we constructed scenarios in which a long cause always produced an effect with a delay sampled from $m_u = 20, i_u = 1$, while a short cause always produced an effect with a delay sampled from $m_u = 10, i_u = 1$. We set $w_c = 1$ and $k_b = 3$ when simulating stimuli. Figure 4.7 demonstrates that both models favored the short cause under the short prior, as indicated by a positive log-likelihood ratio of the short cause over the long cause. Conversely, under the long prior, both models favored the long cause. This tendency becomes more pronounced as the number of data points increases. However, it is worth noting that the tendency to favor the long cause under the long prior is not as strong as the tendency to favor the short cause under the short prior, highlighting the natural advantage of shorter delays.

4.4 Learning from continuous-time evidence

4.4.1 Continuous, effect specified

Lagnado & Speekenbrink (2010) Our first case study revisits the "earthquake" experiment conducted by Lagnado & Speekenbrink (2010). The experiment aimed to investigate the effects of three types of seismic waves (red, yellow, and green) on the occurrence of earthquakes. Unbeknownst to the participants, only one of the three types of waves (referred to as the cause) actually raised the occurrence of earthquakes, while the other two types (referred to as lures) had no effect. This setup allows for the consideration of three structures in the hypothesis space, as shown in Figure 4.8. In each trial, the cause wave occurred 10 times and had a probability of 80% of resulting in an earthquake. The delays between a cause event and its effect event could either be short (3±0.1 s) or long (6±0.1 s), varying across different trials. Meanwhile, the two other lure causes could occur in between with a low probability (35%) or a high probability (65%). Additionally, four earthquake events were sampled at random time points to serve as the base rate. The trials lasted for an average duration of 169±84 s for the short-delay condition and 318±157 s for the long-delay condition. None of these parameters were explicitly disclosed to the participants in the experiment instructions.

Participants were asked to provide both "absolute" and "comparative" ratings for the causal properties of each wave. The "absolute" rating allowed participants to independently rate each

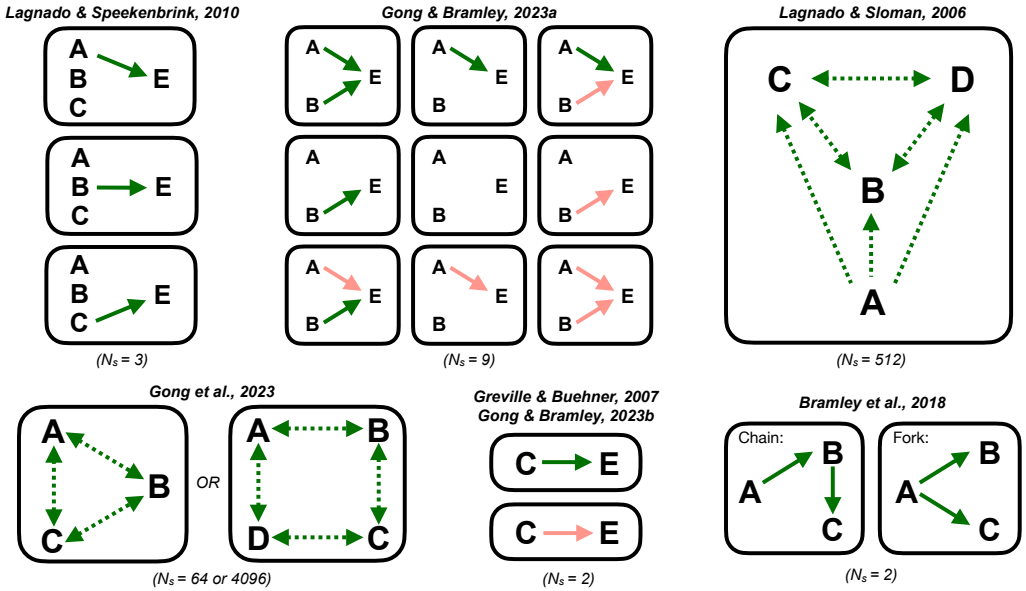


Figure 4.8: The hypothesis space of different studies. Green arrows represent generative links and pink arrows represent preventative links. Dashed arrows $A \to B$ represent two possibilities between two variables A and B: unconnected or $A \to B$, and dashed bidirectional arrows represent four possibilities between two variables A and B: unconnected, $A \to B$, $B \to A$, or $A \leftrightarrow B$. Exogenous links (base rate) are ignored in all graphs.

wave, while the "comparative" rating required participants to allocate ratings for the three waves such that their ratings summed up to 100. Both types of ratings revealed the same pattern: Participants assigned higher ratings to the genuine cause wave compared to the two lures; the rating was influenced by the probability of intervening events but not the delays (see Figure 4.9a). We here model the "comparative" rating. It can be interpreted as a comparison of the probabilities associated with the three causal structures shown in Figure 4.8. In our simulation, we assume the following parameter distributions: $w_c \sim U(0,1)$ for the cause probability (i.e. $\lambda_1 \sim U(0,1)$ for the rate-based model), $\mu_b \sim U(0,100)$ for the base rate mean (i.e. $\lambda_0 \sim 1/U(0,100)$ for the rate-based model), $\mu \sim U(0,100)$ for the cause delay (or influence) mean, and $\sigma^2 \sim U(0,\mu^2)$ for the cause delay (or influence) variance. We generate a Monte Carlo sample of size $m = 10,000$ to approximate the Bayesian inference process. This number will be used for all datasets later on unless parameters are assumed known to the model. [5]

Both event-based and rate-based schemes successfully identified the genuine cause in each condition of the "earthquake" experiment. Moreover, similar to the participants' responses, the

[5]One parameter that could be considered under the rate-based scheme is the time bin configuration. That is, we need to decide the duration considered to constitute one time unit. A cause is presumed to happen in a time bin before the time bins of its effects. We here used 1 second for Lagnado & Speekenbrink (2010); Gong & Bramley (2023a); Gong et al. (2023); Lagnado & Sloman (2006) and 1 day for Greville & Buehner (2007); Gong & Bramley (2023b). Both choices can be regarded as natural. We used 300 milliseconds for Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018) given that more coarse choices would compromise the accuracy.

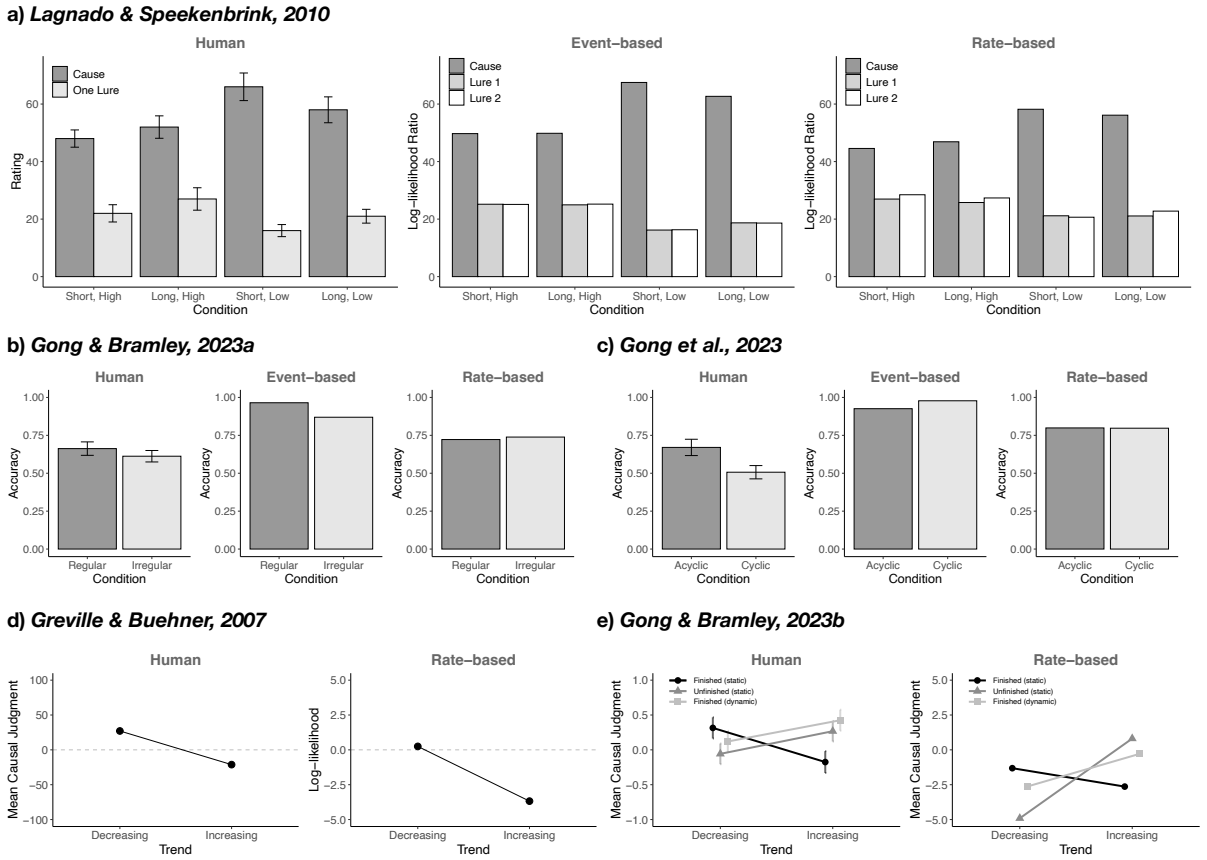


Figure 4.9: Qualitative results of five datasets. A softmax parameter of 10 was applied to Lagnado & Speekenbrink (2010) for visualization. Ratings in Greville & Buehner (2007) are reversed so that they are aligned with Gong & Bramley (2023b) where positive numbers indicated harmful influence and negative numbers indicated beneficial influence.

models were more influenced by the intervention probability than the delay length. This finding addresses the question raised in Chapter 3 that people do not always exhibit a preference for short delays, even based on the uninformative prior. The absence of the short-delay principle in this particular experiment can be attributed to the relationship between the delay lengths and the duration of observations. The long delay was twice as long as the short delay, while the total duration of observations was also twice as long in the long-delay condition. Consequently, the base rate expectation differed between the two conditions, effectively canceling out the effect of delay. We can interpret this as participants mentally defining different durations as one time unit in the two conditions, leading them to arrive at similar conclusions. A similar observation was made in an ecological experiment by Zhang & Rottman (2021a), where no effect of causal delays was found for delays lasting 1 to 9 hours. This lack of effect could similarly be attributed to different participants defining their observation windows differently.

Gong & Bramley (2023a) A similar dataset was collected in Gong & Bramley (2023a, see also Chapter 5). Participants were presented with a causal device consisting of one target component (Effect E) and two control components (Cause A and B). The relationships between each control component and the target component could be generative, non-causal, and preventative, resulting in nine possible causal structures (see Figure 4.8). A generative cause event would always produce an effect event after 1.5±0.5 s. A preventative cause event will cancel any upcoming effect events in the subsequent 3±0.5 s. The effect component can also activate spontaneously. Participants were randomly assigned to the regular base rate or the irregular base rate condition. Each base rate event occurred 5±0.5 s after the previous one in the regular condition, or 5±5 s (according to a memoryless exponential distribution) in the irregular condition. Participants watched the device being intervened on by someone (simulation) for a total duration of 20 s with three interventions on A and three on B.

Given that participants in the study were provided with information about the delay parameters, we make the assumption that our model are also aware of these parameters. Specifically, for generative causes, we set $w_c = 1$ (i.e. $\lambda_1 = 1$ for the rate-based model), the generative delay $\mu = 1.5$ and $\sigma^2 = 0.25$. Regarding the base rate, we assume a mean of $\mu_b = 5$ (i.e. $\lambda_0 = 1/5$ for the rate-based model).

In the case of preventative causes, the event-based scheme assumes the duration of preventative windows follows a gamma distribution $Gamma(\mu_p = 3, \sigma_p^2 = 0.25)$ (Figure 4.10a). All events occurring within the window are assumed to be canceled. The rate-based scheme models the dynamics of preventative influence. It should be noted that the preventative rule employed here does not involve an incubation process for prevention. Instead, the preventative window persists at the maximum level, effectively canceling all effects, for a certain duration. As such, the rate-based scheme captures the preventative causal influence using the gamma cumulative density function, as illustrated in Figure 4.10b, and assumes a maximum level denoted as $\xi = 1$.

The given instructions imply three mechanical rules that can be implemented by the event-based model but not by the rate-based model. Firstly, a single generative cause event only leads to one additional event in the effect component, which is consistent with the setup of the earthquake experiment described earlier. Secondly, in the regular condition, the base rate events occur periodically. Therefore, instead of utilizing a memoryless exponential distribution, the event-based model can employ a gamma distribution $Gamma(\mu_b = 5, \sigma_b^2 = 0.25)$, to model the delay between two consecutive base rate events. In contrast, since the rate-based model does not differentiate between effects generated by base rate events or generative causes, it is unable to leverage the regularity of the base rate and thus treats the regular and irregular conditions in the same manner. The third rule pertains to the preventative window. In the generative process, it is assumed that within a fixed preventative window, all expected effects that are supposed to occur will be canceled, while any expected effects after the window will remain unaffected. Consequently, the size of the preventative window should be inherently smaller than the interval

between a preventative cause event and its nearest effect event E'. The absence of an effect expected to occur after E' can no longer be attributed to prevention. On the other hand, the rate-based model represents prevention as a probabilistic influence, defining a soft window rather than a strict, deterministic window.

As shown in Figure 4.9b, qualitatively, the event-based scheme has higher accuracy compared to the rate-based scheme. It also displays a similar pattern of performing better in the regular condition compared to the irregular condition, which aligns with human performance. In contrast, the rate-based scheme demonstrates a slight tendency to perform better in the irregular condition, potentially attributed to the alignment of the base rate mechanism.

For this dataset and the subsequent datasets, we analyzed two types of correlations between the model and human judgments. The first type is the Pearson correlation, for which we incorporate a softmax parameter to account for the stochastic nature of judgments (Luce, 1959). [6] We utilized a single parameter that was fitted across all conditions for each dataset. The second type is the Spearman correlation, which assesses the ranking agreement between human and model judgments. This provides insight into how well the model captures the human dataset without introducing an additional free parameter. The results are depicted in Figure 4.11a and 4.12a. Both the event-based and rate-based schemes successfully captured human judgments, regardless of whether the conditions were regular or irregular. The event-based model demonstrated slightly superior correlations compared to the rate-based schemes, suggesting that participants may have taken into account at least one of the three mechanistic rules discussed earlier during their reasoning process.

4.4.2 Continuous, effect unspecified

Gong et al. (2023) When the effect variables are left unspecified, the number of potential structures increases quickly. Even when considering only generative relationships, there are four possible relationships between two variables: one-directional, reverse one-directional, bidirectional, and unconnected. Consequently, for three variables, there are 64 potential structures, and for four variables, there are 4096 potential structures (see Figure 4.8). Gong et al. (2023, see also Chapter 6) investigated how individuals navigate and learn from a vast hypothesis space by actively intervening in a causal system. In each causal system, for causally related components, an activated component would probabilistically trigger the activation of each of its effect components once after a delay of 1.5±0.1 s in the regular condition, or 1.5±0.7 s in the irregular condition.

[6] The softmax parameter θ was used to maximize the log-likelihood between models' and participants' choices in Gong & Bramley (2023a); Gong et al. (2023); Lagnado & Sloman (2006), where participants were asked to choose whether each causal connection existed or not. The parameter was used to maximize the linear correlation based on a non-linear transformation $y = sign(x)|x|^{\theta}$ in Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018); Greville & Buehner (2007); Gong & Bramley (2023b) where participants provided ratings for how likely each connection existed or how strong each causal strength was.

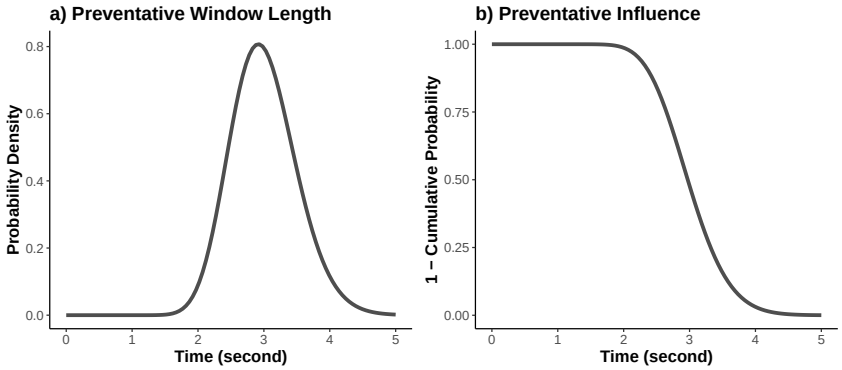


Figure 4.10: The preventive windows and preventative influences. a) The event-based scheme assumes the length of preventative windows are sampled from a gamma density function. b) The rate-based scheme assumes the preventative influence (how much percentage of the effects would be prevented) is relevant to a gamma cumulative function.

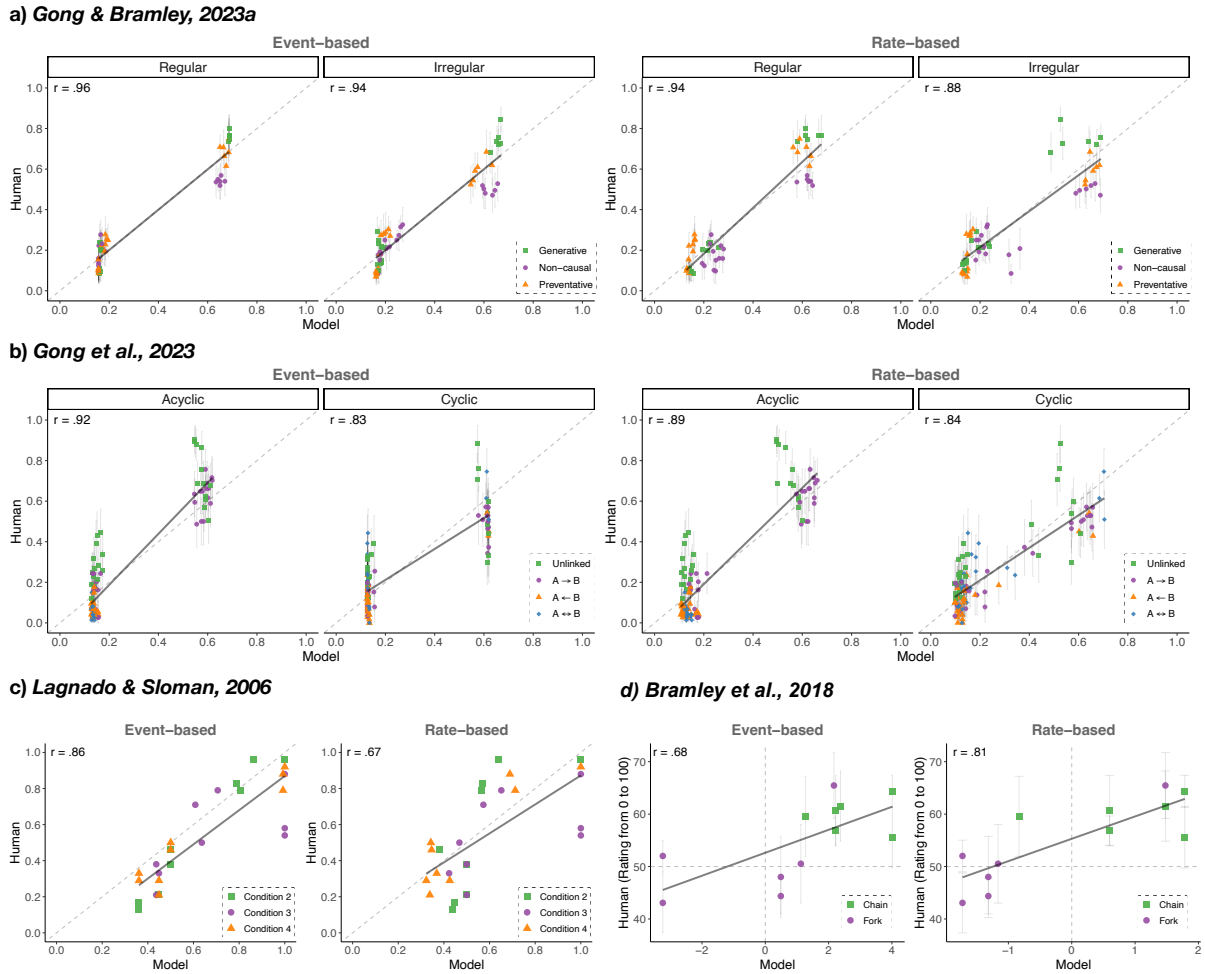


Figure 4.11: The Pearson correlation between model and human judgments. Error bars indicate 95% confidence intervals of human judgments in the dataset whenever the raw data are available.

All causal connections were operational 90% of the time, and none of the components activated spontaneously (i.e. there were no base rate activations). Participants were provided with six opportunities to activate a component in the system during a 45-second interval. Considering the numerous possible connections and the cyclic structures, the number of events recorded in this dataset was significantly higher compared to the aforementioned Gong & Bramley (2023a).

Given that participants in the study were provided with information about the parameters, we assume that models also know the parameters: $w_c = 0.9$ (i.e. $\lambda_1 = 0.9$ for the rate-based model), $\mu = 1.5$, and $\sigma^2 = 0.01$ or $\sigma^2 = 0.49$ depending on the specific regular or irregular condition. No base rate was assumed by both models.

Human results showed a main effect of the structure cyclicity but no main effect of the delay regularity, probably due to that the difference between regular and irregular settings was not pronounced enough (Gong et al., 2023). Therefore, we here focus on the results based on the cyclicity factor alone. In contrast to humans who performed better in the acyclic condition than the cyclic condition, the event-based model demonstrates better performance in the cyclic condition compared to the acyclic condition (Figure 4.9c). It reflects the event-based model is able to leverage the large amount of event information in the cyclic structure (Gong et al., 2023). Conversely, the rate-based model does not demonstrate the same tendency. Due to its limitations in differentiating actual causation, this model fails to leverage the abundance of cyclic events as effectively as the event-based model does.

In terms of both correlation measurements, the event-based model demonstrates better performance in capturing human judgments in acyclic structures, while the rate-based model performs better in capturing human judgments in cyclic structures (see Figure 4.11b and 4.12b). This may suggest that as the number of events increases, the exact computation becomes impractical, necessitating the relaxation of certain constraints within the event scheme to enable more efficient approximations. The event-based model's ability to handle acyclic structures more effectively indicates its advantage in situations where precise computations are feasible. On the other hand, the rate-based model's capability to capture cyclic judgments highlights its ability to approximate across a larger number of events, providing a more efficient approach in such scenarios.

4.4.3 Episodic, effect specified

Greville & Buehner (2007) The episodic evidence could be seen as a combination of contingency and temporal information (Greville & Buehner, 2007). It involves the observation of multiple individuals over a specific time period (Figure 4.2b). So far, research on episodic evidence often focuses on cases each type of event occurs at most once within the observed period. This means that the evidence within each individual's experience may not be very informative. However, by considering multiple cases, the reasoner can compensate for the limited information within each instance and make more informed conclusions about the causal structures.

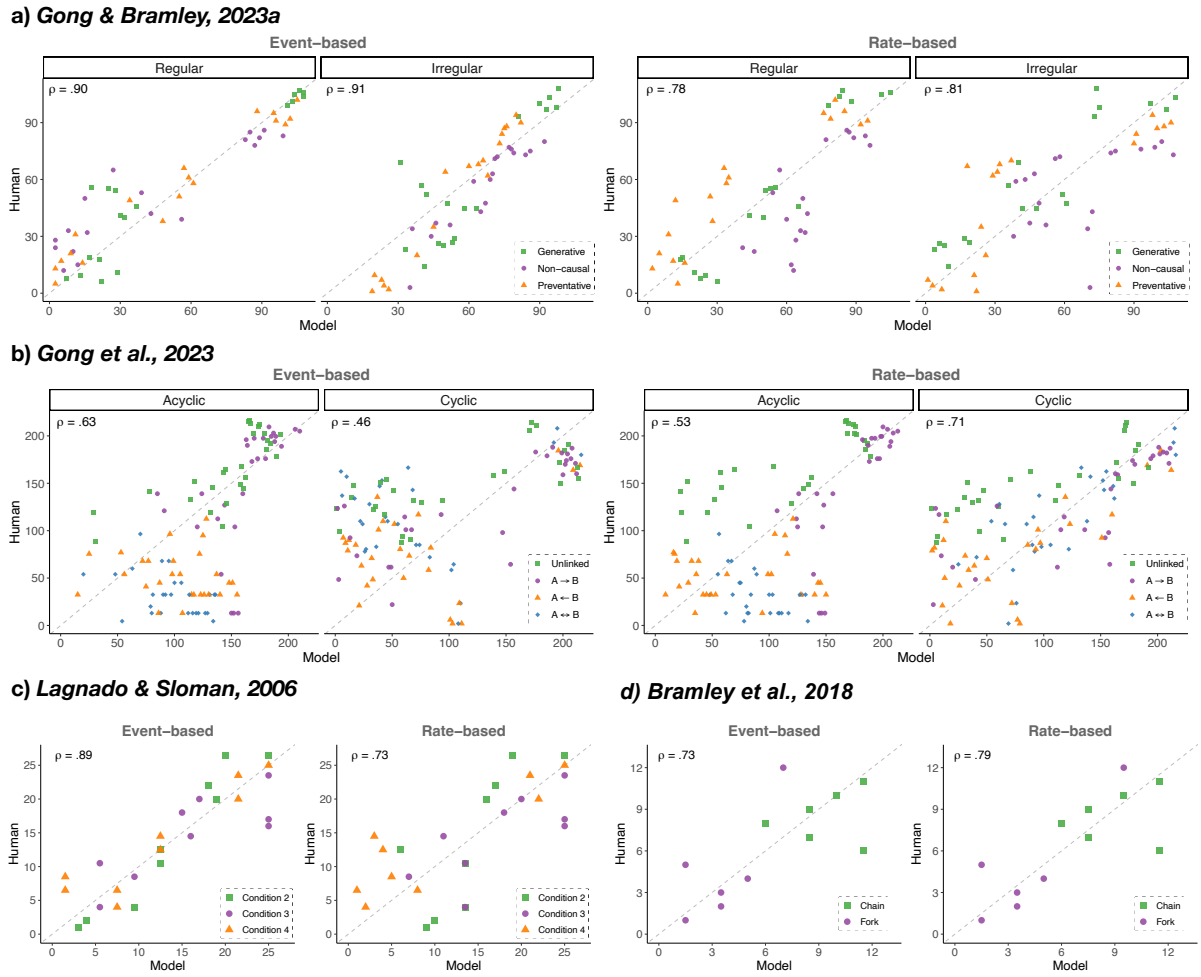


Figure 4.12: The Spearman correlation between model and human judgments.

In Greville & Buehner (2007), participants examined the influence of a ray treatment on the survival of bacterial cultures. Bacterial cultures were randomly assigned to the experimental group, which received a ray treatment at Day 0, or the control group, which did not receive any treatment. Each group consisted of 40 samples. Bacterial cultures were observed from Day 1 to Day 5. The number of new deaths occurring each day was recorded. Participants were asked to rate whether they perceived the treatment as harmful or beneficial based on the observed outcomes in both the experimental and control conditions. Results found, after controlling for the total number of deaths over the 5-day period, participants judged the treatment as more harmful if there were more deaths at the beginning of the observation period (a decreasing trend) which implied that the treatment accelerated the occurrence of deaths. Participants judged the treatment as more beneficial if there were more deaths towards the end of the observation period (an increasing trend) which implied that the treatment delayed the onset of deaths.

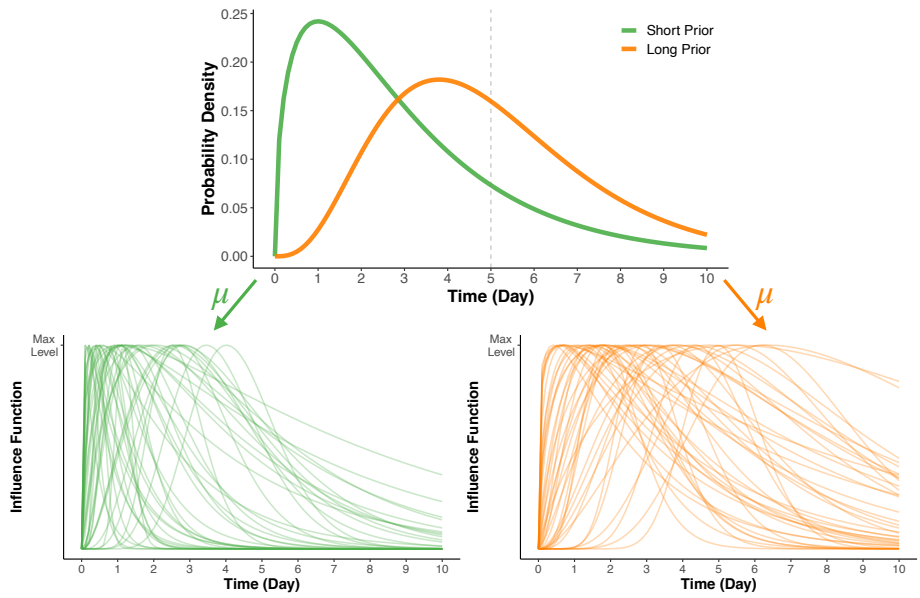


Figure 4.13: The short-delay and long-delay priors regarding the timing of when the cause will take effect on average (Greville & Buehner, 2007; Gong & Bramley, 2023b). The parameter μ is sampled from different prior distributions to form different causal influence functions.

Gong & Bramley (2023b) In contrast to the traditional interpretation based on contiguity (Greville & Buehner, 2007), Gong & Bramley (2023b, see also Chapter 7) proposed an alternative way to interpret the data from the bacterial cultures study. They suggested that if learners rely on the concept of "trend" rather than "contiguity" when making judgments, they may suspect the treatment will ultimately prove harmful if the experimental condition has a worryingly increasing trend. Gong & Bramley (2023b) presented participants with more ambiguous data, where a majority of the forty samples were still alive on Day 5. Participants in the "Unfinished" condition were informed that the observation had not yet concluded, while participants in the "Finished" condition were told that the observation had finished (as Greville & Buehner, 2007). Results in the Finished condition replicated Greville & Buehner (2007). However, in the Unfinished condition, participants interpreted an increasing trend in deaths as indicative of harm caused by the treatment, and a decreasing trend as indicative of benefit (see Figure 4.9e). In a follow-up experiment, Gong & Bramley (2023b) asked the participants to click a button to reveal the data sequentially day-by-day. Under this dynamic display, participants relied on trend rather than contiguity to make causal judgments regardless of the Finished vs. Unfinished instruction framing.

These findings highlight the influence of instructional cues on participants' prior beliefs and the way they interpret the observed data. To model the human judgments, we here assume that instructions tend to influence the learner's prior expectation about causal delays as well the use of data, while visual formats tend to influence the use of data. As shown in Figure 4.13, if participants are informed that the experiment ends at Day 5, they may tend to form a prior

belief that the relevant causal influences would likely to occur within 5 days. When participants were led to believe that the observation had not finished, they anticipated the possibility of longer causal delays. We here assume that, for the Finished instruction (Greville & Buehner, 2007; Gong & Bramley, 2023b), the causal delay (or the expected time of the influential function in the rate-based context) μ is sampled from a gamma distribution with mean of 3 (days) and a variance of 6. For the Unfinished instruction (Gong & Bramley, 2023b), μ is sampled from a gamma distribution with mean of 5 (days) and a variance of 6. The choices here are to make sure that a range of 0 to 5 would cover most of the sampled μ under the Finished instruction (83%) while cover only half of the sampled μ under the Unfinished instruction (57%, Figure 4.13). We assume that under the Unfinished instruction or the dynamic display, participants also used the current trend to extrapolate what would happen in the future, thus treating it as additional evidence. We used a linear regression model to generate data from Day 6 to Day 9 to capture this notion.

The observed death of bacteria cultures in Greville & Buehner (2007) and Gong & Bramley (2023b) could result from either the treatment or the natural death (i.e. the base rate), and it could only happen once for each individual. This situation presents two challenges for the event-based scheme. Firstly, since only one effect event was recorded for each sample, it became difficult for the event-based scheme to track the delays between base rate events without additional assumptions or data reconstructions. In contrast, as shown in Figure 4.3e, the rate-based scheme could simply aggregate samples and assume a constant base rate. Secondly, in this case, multiple underlying mechanisms could be at play. If the treatment was harmful, it could indifferent kill cultures that preempted natural death, or it could hasten the death, the timing of which depended on the expected lifespan (age) of different cultures. A mechanistic approach would require exhaustive consideration of these different mechanisms. Conversely, the rate-based model could bypass the need for detailed mechanistic knowledge, presuming that various mechanisms could result in a similar rate pattern: a rate increase over a certain time period. This higher-level pattern could then be used for making judgments. Consequently, in certain cases, embracing ignorance of the specific mechanisms and relying on rate-based models may show advantageous for reasoning about causation.

Consequently, we focused on the rate-based scheme here and consolidated the data as shown in figure 4.3e. Since the data were collapsed, the rate of how many events happened per day would depend on the total sample size (i.e. forty in both studies). We assume that participants made the judgment by comparing the likelihood between a generative structure and a preventative structure (see Figure 4.8). To model the data, we set $\lambda_0 \sim U(0, 40)$, $\lambda_1 \sim U(0, 40)$ since there were at most forty cases in each group. We set $\xi \sim U(0, 1)$ as the max level of preventative influence (i.e. the beneficial influence). Similar to previous datasets, we set the variance $\sigma^2 \sim U(0, \mu^2)$.

The qualitative results are shown in Figure 4.9d and 4.9e. Participants' inclination in Greville & Buehner (2007) and Gong & Bramley (2023b) was captured by the model. Under the Finished

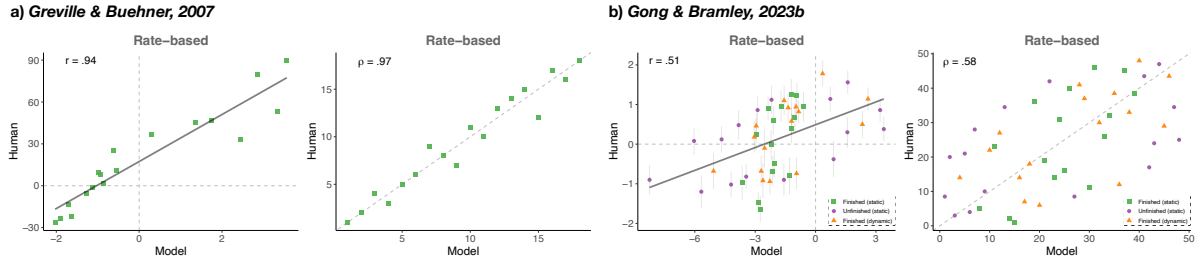


Figure 4.14: The Pearson (r) and Spearman (ρ) correlations between model and human jugdments in Greville & Buehner (2007) and Gong & Bramley (2023b). Error bars indicate 95% confidence intervals of human judgments in Gong & Bramley (2023b).

instruction and the static display (Greville & Buehner, 2007; Gong & Bramley, 2023b), participants and the model both treated decreasing trends as more harmful than increasing trends, showing a contiguity consideration. Under the Unfinished instruction (Gong & Bramley, 2023b), participants and the model both treated increasing trends as more harmful than decreasing trends, showing a trend consideration. Under the Finished instruction and the dynamic display (Gong & Bramley, 2023b), participants and the model demonstrated a trend consideration as well, indicating that with the extrapolated data, the reliance on trends can also show up even under short-delay priors.

Nevertheless, humans demonstrated a general tendency to favor harmful judgments than the model. It can be due to that with a generative (harmful) cause, the ideal learner would expect the experimental condition to consistently exhibit a higher death rate compared to the control group. In Gong & Bramley (2023b), there were cases where the experimental condition had a higher death rate than the control condition during the initial days, but then a lower death rate in the later days even the majority of samples were still alive (e.g. "3, 4, 2, 1, 0" for Day 1 to Day 5 in the experimental condition and "2, 2, 1, 2, 3" in the control condition). In contrast, in Greville & Buehner (2007), the lower number of deaths observed in the experimental condition during the later days (e.g. "16, 12, 8, 4, 0" for Day 1 to Day 5) could be explained as most of the samples may have already died out. Although we accounted for this information in our model and ensured that the expected rate at Day t did not surpass the remaining surviving samples, this inherent sample-size constraint could potentially introduce a bias towards favoring preventive causation.

The quantitative results were shown in Figure 4.14. The rate-based model achieved a good fit with human judgments in the case of Greville & Buehner (2007) and a moderate fit with the judgments in Gong & Bramley (2023b), indicating that participants' causal judgments based on episodic temporal observation are also predictable from a rational perspective. The slightly lower fit in the latter study could be attributed to the more ambiguous nature of the stimuli used in that study.

4.4.4 Episodic, effect unspecified

The final category we consider here is episodic evidence where the effect variable is unspecified. The two datasets both contain evidence that each kind of event can only happen once in each episode. Given that both datasets happen to have no base rate effect, we can model them using both event-based and rate-based schemes.

Lagnado & Sloman (2006) In Lagnado & Sloman (2006), participants were asked to imagine a situation that computer virus can spread through the network and told that the time at which a computer revealed its infection could occur after a variable delay, so later than the time at which the computer became infected. Participants were told that each connection, if existed, worked 80% of the time, and the virus could not reach a computer unless they had been sent from another equipment (e.g. no base rate). Participants watched 100 clips showing the event sequence of when virus appeared in different computers, and were asked to judge the existence of causal links in the system. (see Figure 4.8).

The experiment included four different conditions, but the underlying ground truth structure was consistently: A was the cause of B, and B was the common cause of C and D. This means that in each trial, computers C or D would never become infected without computer B being infected. Since the actual infection time was varied and unknown, the presumed solution is to rely on the conditional probability. However, the timing of virus appearance in each computer could be misleading. For example, in Condition 3 where 50% of trials followed the order of $A - D - C - B$, participants judged the links $A \to D$, $D \to C$, and $C \to B$ were more likely to exist than other links. Their answers cannot explain other trials when only AB, ABC or ABD happened. This suggests that people's reliance on temporal information is so strong that it could not, in this case, be overshadowed by contingency information.

There was a one-second delay between events in subsequent time steps (t_1, t_2, t_3, t_4; see Table 2 in Lagnado & Sloman, 2006). As such, each trial lasted 4 s. We model the dataset using the parameters $w_c = 0.8$ (i.e. $\lambda_1 = 0.8$ for the rate-based model), $\mu \sim U(0, 10)$, $\sigma^2 \sim U(0, \mu^2)$. The base rate is assumed to be zero. In this dataset, the main difference between the event-based and rate-based schemes is that the former address the rule that an event can only occur once for a specific equipment. Consequently, the event-based scheme outperforms the rate-based scheme in accurately capturing human judgments, as shown in Figure 4.11c and Figure 4.12c. Our model successfully explain the phenomenon that temporal information can outweigh contingency information in human causal judgments.

Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018) Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018) tested whether people can differentiate between two causal structures, chain and fork (see Figure 4.8), using solely delay information. Each trial consisted of 12 episodes, wherein events always occurred in the order of $A - B - C$. However, there were variations in the

delay variances between the structures. In the chain structure ($A \rightarrow B \rightarrow C$), the delay variance between B and C was small, whereas the variance between A and C was large, as it encompassed the variability between two stages. Conversely, in the fork structure ($B \leftarrow A \rightarrow C$), the delay variance between A and C was small, while the variance between B and C was large, as there was no direct causal link between the two variables. Participants were asked to judge by distributing 100 percentage points across the two structures.

In contrast to previous datasets, we utilized the "independent delay" parameterization, as described in the original paper (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018), which allowed for distinct delays between different links in the causal structure. It means that to choose between chain and fork, we only need to model the delays between B and C in the chain hypothesis and the delays between A and C in the fork hypothesis. Each episode lasted no more than 3 s. We assume $w_c = 1$ (i.e. $\lambda_1 = 1$ for the rate-based model), $\mu \sim U(0, 10)$, $\sigma^2 \sim U(0, \mu^2)$, and no base rate.

Results are shown in Figure 4.11d and 4.12d. The rate-based model demonstrated a better fit to human judgment compared to the event-based model. The event-based model showed a overall bias towards chains (judging all chain devices as chains and also judging some forks structures as chains). This could be due to that $A-C$ delays (calculated under the fork structure) were always longer than the $B-C$ delays (calculated under the chain structure) in the stimuli. As a result, the preference for the chain structure can be interpreted as an alternative form of favoring the fork structure. In contrast, under the same number of data points, the rate-based model, which assumes macro causal dynamic changes, would have greater tolerance for variations of causal influence more than the event-based model.

We note here that the computational cost of the rate-based model is not always lower than that of the event-based model. In the case of this dataset, the rate-based model may actually be more computationally demanding depending on the chosen time bin configurations. Conversely, the event-based model benefits from the fact that each type of event occurs only once in each episode, resulting in a limited number of causal pathways to consider within each hypothesis.

4.5 General discussion

We develop a rational framework that utilizes the temporal information to make causal inferences, and test people's sensitivity to this rational perspective under different empirical datasets. Our framework expands upon the causal graphical model by incorporating a likelihood function for the calculation of temporal information. This rational framework successfully accounts for three key phenomena discovered in literature: People tend to attribute stronger causality when the delays between cause and effect (1) are short rather than long if expectations are weak; (2) are unvaried from case to case; (3) align with their expectations.

Apart from elucidating people's general rules for thinking about causal delays, a bigger question is how to utilize temporal information in more complex, continuous-time evidence to uncover the underlying causal structure. Our framework can handle with a range of temporal causal learning tasks spanning from 2006 to the present. These tasks encompass various scenarios, such as continuous evidence revealed on a single timeline or episodic evidence with independent shorter timelines. They also vary in the hypothesis space size, the inclusion of background, preventative, or cyclic causes, as well as the instruction details regarding delay expectations. We show a level of consistency between judgments from the rational framework and judgments from people. Thus, people not only are capable of utilizing temporal information in diverse causal learning situations but also reveal a systematic, predictable, and to some extent, rational pattern in their judgments.

4.5.1 A pluralistic view

We have presented two schemes (event-based and rate-based) under a unified Poisson-Gamma framework. The existence of a pluralistic view is not a new concept in the field of causal cognition. In the realm of causal attribution, where individuals are asked to make judgments regarding specific events, researchers have debated the relative importance of covariation versus process (Gerstenberg et al., 2021; Sloman, 2005; Wolff, 2007; Lombrozo, 2010). The question arises whether people prioritize imagining how the outcome would have changed if the cause had been different (Sloman, 2005; Icard et al., 2017), or if they focus more on determining if there was physical contact between the cause and effect (Wolff, 2007; Talmy, 1988). Instead of exclusively relying on a single level of abstraction, individuals demonstrate a pluralistic view by considering both the *occurrence* of the outcome and the *manner* in which it occurred given the state of the cause (Gerstenberg et al., 2021). We next demonstrate two reasons why a pluralistic perspective is also important in the domain of temporal causal learning, from the concerns of mechanisms and computational costs, respectively.

For a concern of mechanisms

Learning a causal system could be based on different type of evidence: atemporal, temporal, and spatiotemporal. As we move from atemporal to spatiotemporal evidence, the richness of information increases. Atemporal evidence contains a limited number of categories. Samples from the same category (e.g. a category where the evaluated cause is present while the outcome effect is absent) are often treated as identical, as there is no additional information to distinguish them (Allan, 1980; Cheng, 1997; Perales & Shanks, 2007; Griffiths & Tenenbaum, 2005). Researchers also examine spatiotemporal evidence when asking individuals to make causal inferences for 2D physical scenes (Ullman et al., 2018; Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018; Gerstenberg et al., 2021). Due to the brevity of the clips used in these studies, it is crucial to leverage the

mechanistic theory to discover the underlying structure. This necessitates that learners consider the specific movements of objects that are causally related to one another.

We argue that temporal evidence shares characteristics with both atemporal and spatiotemporal evidence. Like atemporal data, temporal evidence can provide multiple samples, as effect events may occur multiple times without the necessity of having individual identifications (e.g. the Melatonin example before; see also Pacer & Griffiths, 2012; Griffiths & Tenenbaum, 2005). It allows for type-level reasoning, where we can reason about how the rate of effect occurrence changes after a putative cause occurs. At the same time, temporal information can also enable token-level thinking. When one cause produces a very limited number of effect events (e.g. one on an effect entity), by taking into account delays between cause and effect, and the prior knowledge of causal delays, we might infer which specific occurrence of the cause was responsible for this specific occurrence of the effect. As such, our framework allows us to deal with different situations accordingly to what the mechanistic or ontological commitments match with.

For a concern of computational costs

Continuous time allows for precise temporal information, with each event having its unique time point and relationship with all other events. Events of different entities are often intermingled, and events on the same entity may happen multiple times within the same observing episode. However, this precision and combinatorial credit assignment issue poses challenges as considering process would result in an infinite number of possibilities. Moreover, when observing a causal system in continuous time, any event occurring in the present moment could theoretically be a result of any event that happened in the past. For instance, an event happening to a person in their forties may be greatly influenced by a decision they made in their twenties, a decision they may not have regarded as significant. Given this richness of potential relationships, how can human beings manage such complexity without considering every single possibility? We propose that a way to manage complexity is by focusing on the macro level of how the rate of effect changes. This approach proves useful especially when dealing with a large number of effects. However, determining the appropriate granularity for rate calculation introduces another perspective that needs to be explored. Nonetheless, it is important to note that reasoning from temporal evidence introduces a more significant computational challenge compared to the previously studied atemporal learning setting. Therefore, it could serve as a good learning setting to study bounded rationality (Simon, 1982; Lieder & Griffiths, 2020).

4.6 Conclusion

We inevitably live in a world where events unfold in a continuous temporal manner. In order to learn causal structures from this dynamic world, it is essential to effectively and efficiently

process temporal information. Despite fruitful empirical findings regarding how individuals process temporal information in various causal learning tasks, there is a lack of a unified theoretical framework to integrate these findings. We present a rational framework for causal induction in the temporal domain. This framework can provide predictions of causal judgments for a variety of temporal causal learning task. We show that human performance is sensitive to the rational perspective. Qualitatively, the model demonstrates the phenomena governing causal delays that have been observed in empirical studies. Quantitatively, human judgments exhibit strong correlations with model judgments across different datasets. By establishing a rational framework, we take the initial step towards investigating the underlying mental processes involved in temporal causal learning. This framework serves as a benchmark for further exploration of how humans learn causal structures when faced with limited cognitive resources.

Box 4.1: The desirable properties of gamma distributions for studying continuous time causal learning.

To begin with, gamma distributions have a convenient summing property. If $X, Y \sim \text{Gamma}(\alpha, \beta)$ then $X + Y \sim \text{Gamma}(2\alpha, \beta)$. As an example, suppose Bus #176 arrives every 12 ± 2 min. If you arrive at the bus stop just as it leaves you might expect to wait $\text{Gamma}(\alpha : 36, \beta : 3, [\mu : 12, \sigma : 2])$ minutes. However, if you then are told the next bus is canceled, the expected waiting time will double: $\text{Gamma}(\alpha : 36, \beta : 3) + \text{Gamma}(\alpha : 36, \beta : 3)$ is equal to $\text{Gamma}(\alpha : 72, \beta : 3, [\mu : 24, \sigma : 2.8])$. This neat transition among gamma distributions will facilitate the calculation of *preventative* causation.

The second desirable property is the memoryless property of exponential distribution, a special kind of gamma distributions when $\alpha = 1$. The memoryless property means that the expected delay is constant, no matter how long you have already waited for. The detailed mathematical proof is provided in Box 4.2. In causal learning, we not only encounter regular causal-effect delays, but also irregular delays such as delays between agents' interventions. Gamma distribution family enables us to represent both memory and memoryless time intervals under the same probability distribution group.

The final theoretical consideration comes from the relationship between Gamma distributions and Poisson processes (Pacer, 2016). Event generations follow Poisson processes if the waiting time follows exponential distributions. More importantly, according to the property of Gamma distributions, if we count every 10 events, the waiting time will shift from unreliable $Gamma(1, \beta)$ to reliable $Gamma(10, \beta)$. If we count every 1000 events, the waiting time would follow $Gamma(1000, \beta)$ which is very reliable. Indeed, that is the exact working mechanism of atomic clocks, where we count every 9×10^9 caesium hyperfine transition events as 1 second. Although it is not clear whether biological clocks work

similarly (Buonomano, 2017), representing temporal dynamics as Gamma distribution and Poisson process may help build a bridge between micro and macro perspectives of causal learning in time.

Box 4.2: The memoryless property of exponential distributions. The exponential distribution is one kind of probability distributions with random variables valued in $[0, \infty)$. It includes one rate parameter $\lambda \in [0, \infty)$ to illustrate how many events are expected to happen per united time. It could be seen as a special case of gamma distribution when the shape parameter $\alpha = 1$. Accordingly, its probability density function is:

$$P(T = t) = \lambda e^{-\lambda t} \tag{4.10}$$

The cumulative distribution function is:

$$P(T \leqslant t) = 1 - e^{-\lambda t} \tag{4.11}$$

Therefore, we know that: Exponential distributions are *memoryless*, which means that if we want to see an event t minutes later, but we already wait for s minutes and no event have not happened yet, the probability of waiting for at least another t minutes would be the same as when we just begin to wait, which means $P(T > s + t|T > s) = P(T > t)$. This can be proved as follow:

$$P(T > s + t|T > s) = \frac{P(T > s + t, T > s)}{P(T > s)} = \frac{P(T > s + t)}{P(T > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} \tag{4.12}$$

$$P(T > t) = 1 - P(T \leqslant t) = e^{-\lambda t} \tag{4.13}$$

Chapter 5

Learning generative and preventative structures in continuous time

Causal cognition studies have largely focused on learning and reasoning about contingency data, but this could just represent the tip of the causal cognition iceberg. A more general problem lurking beneath is that of learning the latent causal structure that connects events as they unfold in continuous time. In Chapter 4, I have shown a normative model for how time information *should* be utilized. In this chapter, I will collect empirical data and investigate how people *can* utilize time information to learn causal structures.

Fewer studies have examined learning and reasoning about systems exhibiting events that unfold in continuous time. Of these, none have yet explored learning about preventative causal influences. How do people use temporal information to infer which components of a causal system are generating or preventing activity of other components? In what ways do generative and preventative causes interact in shaping the behavior of causal mechanisms and their learnability? I explore human causal structure learning within a space of hypotheses that combine generative and preventative causal relationships. Participants observe the behavior of causal devices as they are perturbed by fixed interventions and subject to either regular or irregular spontaneous activations. Participants are capable learners in this setting, successfully identifying the large majority of generative, preventative and non-causal relationships but making certain attribution errors. I propose a family of more cognitively plausible algorithmic approximations. Participants' judgment patterns can be both qualitatively and quantitatively captured by a model that approximates normative inference via a simulation and summary statistics scheme based on structurally local computation using temporally local evidence.

Content from this chapter is a reprint of the material as it appears in Gong & Bramley (2023a).

5.1 Introduction

We naturally think about the world in terms of a progression of events that cause and affect one another. When successful, causal reasoning helps us abstract from our real-time experience to recognize stable causal mechanisms that we can use to explain, predict and sometimes control our environment (Sloman, 2005). However, inferring causal structure in real environments is notoriously challenging, involving a complex interplay between incoming evidence, action, and intuitive theories of how causal influences manifest and link elements of experience like events, objects and variables (Lagnado, 2011; Goodman et al., 2011; Griffiths & Tenenbaum, 2009).

Two of the basic and well-studied notions of causality are generative and preventative relationships. In a generative relationship, we think of the occurrence of one event as bringing about the occurrence of another. A *generative* causal claim implies the counterfactual that, had the cause not occurred, the effect would not have occurred either. In probabilistic accounts of causal reasoning, generative causality is typically linked with an expectation of positive contingency: The presence of a generative causal variable is associated with an increase in the probability of its effect(s) being present compared to cases where the cause is absent or inactive. The reverse of this is the notion of a *preventative* causal relationship, where we think the occurrence of a causal event as blocking another event from occurring. A preventative causal claim implies the counterfactual that, had the cause not occurred, the effect would have occurred. Probabilistically, we thus expect the presence of a preventative cause to decrease the probability of its effect(s) occurring, compared with cases where the cause is absent or inactive (Cheng, 1997; Sloman, 2005; Griffiths & Tenenbaum, 2005).

The majority of causal learning research has focused on inferences from atemporal evidence, which can be represented in tables of co-occurance or contingency that reflect the statistical dependencies among a set of variables (Cheng, 1997; Buehner et al., 2003; Griffiths & Tenenbaum, 2005; Rottman & Hastie, 2014; Lagnado & Sloman, 2004). This kind of data is central in scientific experimentation, in that it depends on the collection of multiple independent samples (Pearl, 2000; Pearl & Mackenzie, 2018; Zimmerman, 2007). However, an intriguing question regarding human cognition is about how people learn causal relationships from temporal data, given that we experience the world as one continuous timeline, and that real world causal mechanisms often take time to produce their effects. The temporal setting also allows that multiple events of the same type may occur multiple times to a single individual. This more closely resembles repeated-measure data from a single individual than reasoning from large independent samples. In this setting, people might rely more on "soft" cues (e.g. time, prior knowledge) than the contingency principle (Lagnado et al., 2007). Understanding how people learn from temporal data is crucial because it not only improves our understanding of the basic mechanisms of human learning, but also clarifies the differences between scientific practices and intuitive causal inference.

Besides this, studies of atemporal causal learning (Cheng, 1997; Buehner et al., 2003; Griffiths & Tenenbaum, 2005; Rottman & Hastie, 2014; Lagnado & Sloman, 2004) as well as recent studies of temporal causal learning (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Buehner & McGregor, 2006) typically focus on one type of causal relationship at a time. In contrast, this paper aims to investigate how one can learn preventative and generative relationships where both are in play at once. Can people identify what is causing and what is preventing an effect *despite*, and perhaps even *because* of the ways that such causal influences intertwine and interact in time. Although this may sound like a "niche" scenario, it is actually very common. To illustrate such an everyday situation: Suppose you adopt a cat that, while adorable, frequently urinates outside its litter. You would like to understand why and learn to prevent this behavior before she completely ruins your soft furnishings.[1] Identifying the causes of the problem peeing, not to mention an effective pee-prevention strategy is nontrivial and might require considerable thought and experimentation. Perhaps you notice the cat rarely pees inappropriately when playing with its teaser. However it is unclear if teaser is an effective preventer, because the times of day she is encouraged to play with it may be different from those when she pees. Intuitively, diagnosis becomes easier if you can exploit the moments when you know she tends to urinate to test whether the teaser is an effective preventer. For instance, if she often urinates around 7 a.m, you could try introducing her teaser around this time. Alternatively, you might consider encouraging her to drink water to stimulate additional need to urinate a little before the time she more habitually plays with her teaser. In this way you might leverage either an established baseline expectation or an established generative cause (extra water) to facilitate your preventative investigation.

The example above shows, firstly, that temporal expectations are necessary to make sensible causal inferences (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Buehner & McGregor, 2006; Greville & Buehner, 2010; Lagnado & Sloman, 2006). In this case we need some sense of when the cat usually pees inappropriately, as well as an expectation of how long it takes for water to pass through its body. Secondly, it is likely that generative and preventative influences *interact* in terms of how they reveal or obscure one another (Rottman, 2016; Lombrozo, 2010). The existence of either a regular base rate occurrence of an effect, or of effects generated by a known generative cause with regular delays, makes it possible to form a strong expectation against which we can test preventative causes.

In this paper we distill these reasoning patterns into a task and a rational analysis that aim to examine: (1) whether people can use temporal knowledge to learn causal systems that include both generative and preventative causes, (2) how the regularity or predictability of the base rate occurrence of an effect of inference affects the learning process, and (3) whether there are interactions between learning different types of causes.

Apart from establishing what factors influence temporal causal learning, we also want to know *how* people learn, i.e. what kind of inference process can capture human judgments. Causal

[1] This is a real life example for one of the authors of this paper.

Bayesian Networks (CBNs) are an established mathematical framework representing and reasoning about causal structure giving rise to observations (Pearl, 2000; Rottman & Hastie, 2014; Allan, 1980). In the psychology of causal reasoning, they have served as a computational-level norm (Marr, 1982) allowing researchers to investigate how the cognitive processes of causal induction approximate or deviate from ideally reverse engineering the generative causal mechanism most likely to be responsible for one's observations. Accordingly, a number of process-level models have been proposed (Bramley, Dayan, et al., 2017; Davis & Rehder, 2020) that each capture some of the ways human performance departs from this kind of Bayesian ideal. However, CBNs and extant process-level models do not describe the role of continuous-time information in human causal structure induction. This is surprising, since as argued, time is a ubiquitous feature of human interactions with their environment, and the need to process rich temporal information in real time is a practical constraint on most of our basic causal inferences. In this paper we take a rational analysis approach (Anderson, 1990; Simon, 1982), starting with a normative account of inference from observations of real-time events to their underlying causal structure and developing a process-level approximation family that can capture human deviations from this. For our normative account, we expand the CBNs framework so that it incorporates representing and learning via causal delay information. Alongside this, we propose a process-level framework that exploits several tricks for approximating intractable probabilistic inference: mental simulation (Ullman et al., 2018; Battaglia et al., 2013), local computations (Bramley, Dayan, et al., 2017; Fernbach & Sloman, 2009), and temporally local evidence (Bramley et al., 2015; Bramley, Dayan, et al., 2017; Bonawitz et al., 2014).

5.2 Question 1: How do beliefs about causal orders and delays shape causal structure learning?

One of our main goals is to test whether people can use their knowledge about time and causality to learn causal structure. Previous studies have demonstrated the temporal knowledge from three perspectives: order, delay expectation, and delay variation.

Foundational to the notion of causation, is the principle that causes must precede their effects (Hume, 1740). Accordingly, people use the *order* of occurrence to constrain and sometimes fully attribute causal structure among components of a system (Bramley et al., 2014). Indeed, event order appears to be a strong heuristic cue to causal order, having been shown to override contingency patterns even in settings where participants are instructed that order is an unreliable guide (Lagnado & Sloman, 2006) or even completely irrelevant to causal structure (Rottman & Keil, 2012).

As well as order, causal inferences are sensitive to *delays* between events. People make stronger or more confident (generative) causal attributions connecting events separated by short temporal delays than by long temporal delays (Shanks et al., 1989; Tarpy & Sawabini, 1974; Shanks & Dickinson, 1991). This reflects one of the most basic forms of learning, in which animals associate stimuli at a learning rate inversely related to their separation in time (Grice, 1948). However, going beyond automatic associations in time, human causal attributions are moderated by domain-specific delay expectations, with shorter-than-expected delays also reducing the causal judgment strength (Buehner & May, 2002; Mendelson & Shultz, 1976; Hagmayer & Waldmann, 2002; Lagnado & Speekenbrink, 2010; Buehner & McGregor, 2006). For example, Hagmayer & Waldmann (2002) found participants judged whether an insecticide prevents mosquitoes by comparing prevalence of mosquitoes in fields with and without the insecticide, but judged whether planting flowers prevents mosquitoes based on whether the prevalence of mosquitoes was affected the year after the flowers were planted, presumably expecting that flowers would take longer to influence the insect population than insecticide. Besides the length of inter-event delays, people are also sensitive to *delay variability* when they are repeatedly exposed to putative cause–effect pairs. That is, people rate one kind of event as less of a strong cause of another to the extent that the delay varies a lot across instances (Lagnado & Speekenbrink, 2010; Greville & Buehner, 2010).

Recently several studies proposed models to capture human's expectations for delay length and variation, including scenarios of pairwise causal learning (Pacer & Griffiths, 2012), structure learning (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Pacer & Griffiths, 2015), imputing hidden causes (Valentin et al., 2022), or making judgments of actual causation given a known causal structure (Stephan et al., 2020). Nevertheless, these studies have predominantly focused on cases of generative causal influence. Additionally, they have focused on inference from sets of independent clips, in which root components are usually activated at the start and effects follow from this. However, a more naturalistic and challenging setting is one where causes and effects intermingle and components can exhibit multiple activations, and both generative and preventative influences can occur within a single learning episode. This is the setting we will explore.

5.3 Question 2: How do generation, prevention, and background causes interact in affecting causal learning?

Early studies of causal cognition focused on elemental pairwise causal judgments based on contingency data. While not directly related to the current temporal setting, these studies reveal

general principles of causal inference. For instance, the ΔP principle captures the change in the probability of an effect occurrence with vs. without a putative cause ($P(E|C) - P(E|\neg C)$), forming a basic metric for the strength and direction of a potential causal effect (Allan, 1980). However, researchers later found people are sensitive to the *base rate* of the effect $P(E|\neg C)$. That is, how frequently the effect occurs in the absence of the cause. For a fixed ΔP, people infer stronger generative influences when base rates are high (because this implies the cause would have succeeded a greater proportion of the time if it had the chance to operate), and stronger preventative influences when base rates are low (Buehner et al., 2003; Cheng, 1997; Wu & Cheng, 1999).

In addition to the size of the base rate, the regularity of the base rate also influences causal inference. In Rottman (2016), participants were asked to evaluate the effectiveness of two medications. In one context, the baseline pain level was random from case to case, whereas in another setting, it was autocorrelated (i.e. it tended to increase or decrease smoothly over time). Participants were found to focus more on the raw effect values in the random condition, while focusing more on the change of effect values in the autocorrelated condition. This indicates that people are sensitive to environmental *stability*, adapting how they accumulate and represent causal effect evidence when receiving information in different environments (Biele et al., 2009; Whittle, 1988). We will explore whether people are sensitive to temporal regularity (periodic vs. unpredictable) and, if so, whether or not they adjust their inference strategy accordingly.

Finally, humans show some ability to condition on other variables when inferring the role of a target variable (Rescorla & Wagner, 1972; Gopnik et al., 2001; Beckers et al., 2005; Shanks, 1985). People can use information regarding known causes to better understand unknown causes, particularly preventative causes. The classic paradigm in prevention learning is to let learners build a generative impression of a cause ($A+$), and then expose them to negative results under the combination of a generative cause and a preventative cause ($AB-$). People learn the preventive cause better in this case than when the preventative cause is paired with the negative result alone ($B-$, Melchers et al., 2006; Rescorla & Wagner, 1972; Lovibond & Lee, 2021; Lee & Lovibond, 2021). However, the existence of temporal information may actually increase the difficulty of thinking about causal interactions: To utilize the generative causes to learn about prevention, the learner must have ensured that generative causes would have produced effects in a particular time period when preventative causes are active.

Recent studies also demonstrate human limitations in dealing globally with joint probability, i.e. reasoning probabilistically about multiple interacting variables (Bonawitz et al., 2014; Fernbach & Sloman, 2009; Markant et al., 2016; Griffiths et al., 2015; Davis et al., 2020). Outside of very simple learning problems, they may rather focus on local components of the system rather than maintain a global perspective. For example, people often infer an erroneous $A \rightarrow C$ link when reasoning about a generative system with two links $A \rightarrow B \rightarrow C$, apparently failing to notice that B can explain C's dependence on A (Fernbach & Sloman, 2009; Davis et al., 2020).

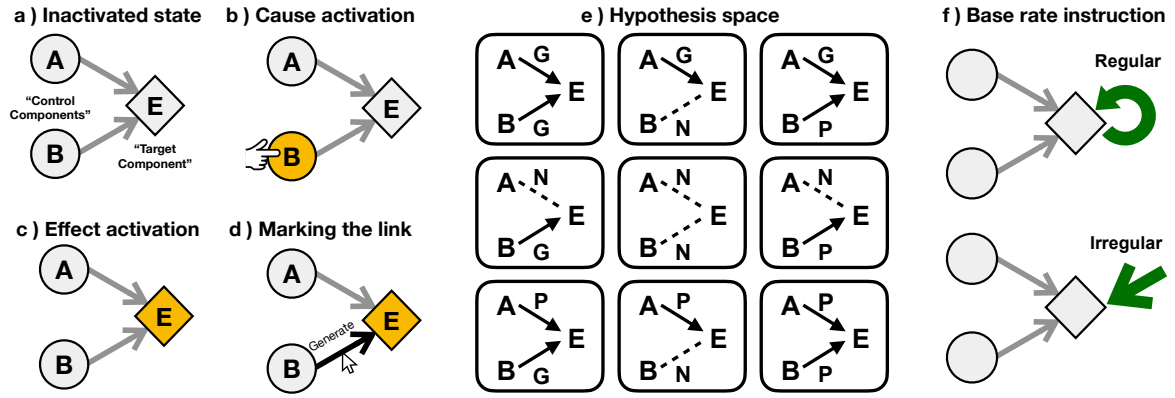


Figure 5.1: Causal devices tested in this paper. a-d) Experimental interfaces. Participants were instructed to the control components and target components in the causal devices and observed how the system reacted to pre-set interventions. They marked their answers of the role of each connection during or after the observation. e) The response hypothesis space (all possible pairwise combinations of generative (G), non-causal (N), and preventative (P) connections). f) The illustrations shown to participants in the regular (periodic) vs. irregular (exogenous) base rate condition.

Through model comparison, we will explore to what extent people can reason globally or locally about causal structure on the basis of real time evidence, e.g. whether they can account for and potentially bootstrap their inferences by considering interactions between causal mechanisms, or if they rather fail to make these accommodations.

5.4 Question 3: How do people process temporal dynamics to make causal inferences?

We build two models for describing how the temporal information could be processed in order to make causal inferences. We will explain the models at a theoretical level in this section and refer the readers to Appendix A.1 and A.2 for technical details. To do this, we first introduce the learning task before describing our model so that readers can get a concrete understanding of how it works.

5.4.1 The learning task

In this study, participants must guess the structure of abstract causal "devices" (Bramley, Dayan, et al., 2017; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Gong et al., 2023) composed of three components (Figure 5.1a–5.1d): two "control components" (i.e. Cause A, B) and one "target component" (i.e. Effect E) on the basis of observations of those structures being perturbed by interventions. To control the impact of interventions, our experiments focus on a learning setting

wherein the interventions are part of the stimuli, meaning participants observe them taking place rather than selecting and performing them themselves.

For each device, the connection between each control component and the target component could be generative, preventative, or they might be unconnected (non-causal). Thus, we focus on learning in a nominal hypothesis space of 9 possible structures including all combinations of generative, preventative and non-causal connections from A and B to E (Figure 5.1e). As a first foray into preventative causation in real-time causal structure induction, we focus on this restricted hypothesis space of causal structures which only contains the common effect topology. However, the experimental paradigm and computational models we introduce can generalize directly to learning in arbitrarily broader causal hypothesis spaces, as well as under different prior expectations about plausible delays and relations.

We focus on relationships between *point events* (i.e. activations) occurring at a device's components at particular moments in time. We assume an activation of a generative component will always produce an "extra" activation of the target component (i.e. causal strength = 1, Cheng, 1997, see Figure 5.2a). We use the gamma distribution to model and generate the delays between causes and effects (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Stephan et al., 2020; Valentin et al., 2022).

We assume an activation of a preventative component blocks any activations of the target component for a short stochastic time window (Figure 5.2b). We assume that prevention occurs irrespective of whether activations would have been caused by a generative causal influence or would have occurred spontaneously. Preventative influences are thus conceived as having a broad preventative scope (C. D. Carroll & Cheng, 2009).[2] By definition, activations of non-causal components have no impact on the behavior of the target component.

Two forms of background activation are considered. In the *Regular base rate* condition, the target component activates quasi-periodically (Figure 5.2c). In the *Irregular base rate* condition it occurs exactly as often overall but is completely unpredictable when the next occurance will be (Figure 5.2d).

5.4.2 Bayesian inference

We now lay out an ideal Bayesian model as a normative model for this task. The ideal reasoner is presumed to take all activation events within the observation interval as the basis of their inference and use the relative likelihood of these under different structural hypotheses to update a distribution over causal structures. The calculation of likelihood here depends on an expensive enumerative actual causal attribution step (Halpern, 2016). The basic idea is that accurate judgments about *type-level* causal relationships (i.e. about the underlying causal structure) depend

[2]We recognize that there are other ways in which one might operationalize prevention and we consider several alternatives in the General Discussion.

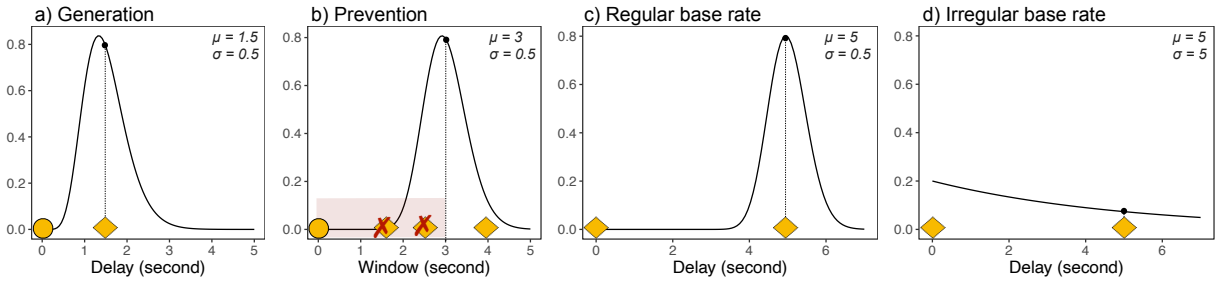


Figure 5.2: Using gamma density distributions to generate the delays between cause and effect and the blocking windows of preventative causes. Circles indicate cause events and diamonds indicate effect events. Each vertical line shows an actual sampled situation. (a) The distribution of delays between cause and effect. When a generative cause event occurs, it will produce an effect event after 1.5 ± 0.5 s. (b) The distribution for preventative window length. When a preventative cause event occurs, all effect events supposed to occur within 3 ± 0.5 s will be canceled, while effects outside the preventative window (the red box) would not be affected. (c) The distribution of delays between base rate events in the regular condition. When a base rate effect occurs, the next base rate effect will occur after 5 ± 0.5 s. (d) The distribution of delays between base rate events in the irregular condition. When a base rate effect occurs, the next base rate effect will occur after 5 ± 5 s.

on detailed considerations about the *token-level* causation giving rise to the observable evidence (i.e. which particular event actually caused which particular effect). There are often a very large number of possible ways that even a single causal hypothesis could have produced a particular pattern of observed events. For instance, if A activates at 0.1 s and B activates at 1.2 s ($\mathbf{i}\{i_A^{(1)} = 0.1s, i_B^{(1)} = 1.2s\}$), and the learner observes two subsequent effects ($\mathbf{d}\{d^{(1)} = 1.5s, d^{(2)} = 2.8s\}$), even under the hypothesis that A and B are both generative causes, the data could be produced in multiple ways: A could have caused the first effect and B the later one ($i_A^{(1)} \rightarrow d^{(1)}$, $i_B^{(1)} \rightarrow d^{(2)}$) or A could have caused the later effect and B the earlier one ($i_A^{(1)} \rightarrow d^{(2)}$,$i_B^{(1)} \rightarrow d^{(1)}$). Alternatively one or both connections could have not revealed their effects yet and meaning either or both observed effects could simply be base rate activations. Therefore, in order to maintain rational beliefs about causal structure, the ideal reasoner considers all possible causal paths that could describe what actually happened given each possible structural hypothesis.

Figure 5.3a shows two examples of the tree of possible causal paths under two of the possible structural hypotheses. Since one must consider possible causal paths exhaustively, the complexity of this inference scheme scales in a worse-than-polynomial manner as the number of events a learner observes increases.

5.4.3 Simulation-and-summary-statistic approximation

While the enumerative approach achieves benchmark performance by inverting the generative model, exhaustively considering pathways linking all observed events, it makes unrealistic demands on memory storage and computing power compared to what could plausibly be involved in human cognition. Therefore, we propose a process-level model that is more consistent with cognitive

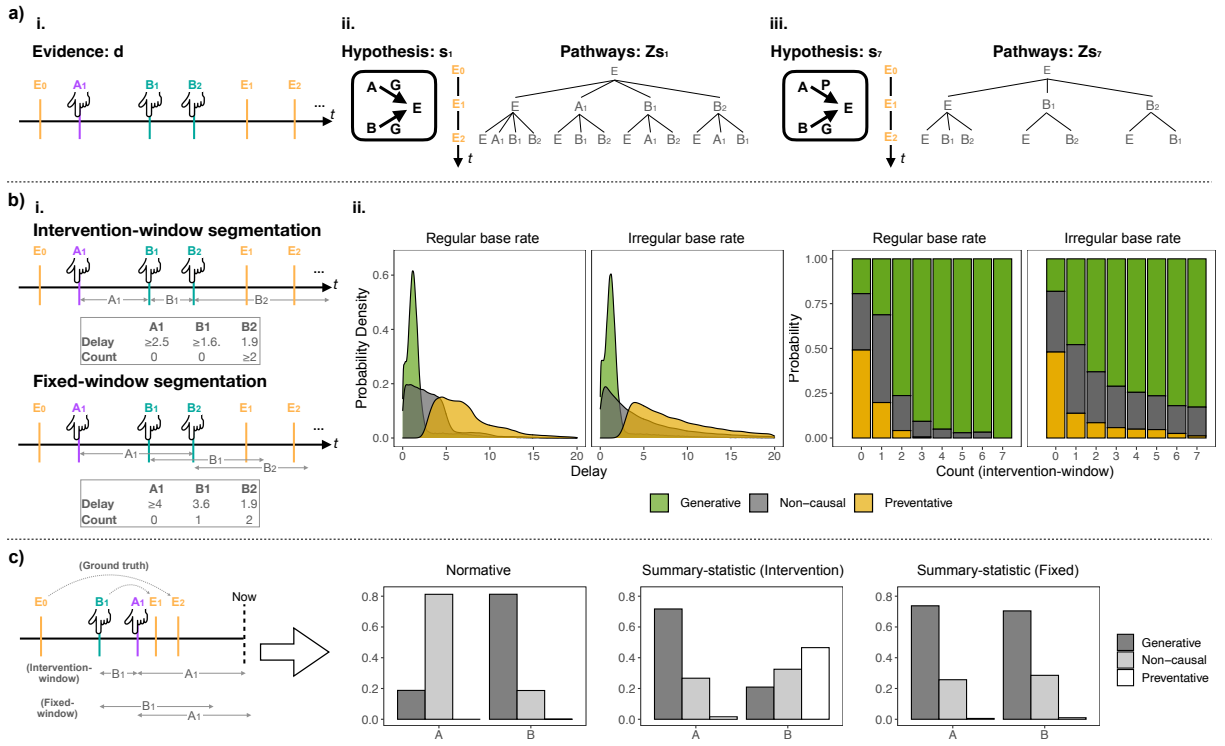


Figure 5.3: Illustrations of model algorithms. a) Causal path construction under fully normative inference. i. Data: Each line indicates a cause (A, B) or an effect (E) event in the evidence. ii-iii. Ideal observer sums over all possible pathways (branches) that explain all events evidence under each hypothetical structure. ii. e.g. Under the structure where A and B are both generative causes, there are 13 ways to explain Evidence d: one candidate cause for E_0 (base rate), four candidate causes for E_1, and 3–4 candidate causes for E_2 depending on how E_1 is explained. iii. Possible pathways under a different structure. b) Summary-statistic approach: i. Intervention-window or fixed-window evidence segmentation. ii. Distributions for summary-statistics given different connection types based on pre-simulated data. The model uses likelihood of observed statistics under these distributions as a proxy for generative model likelihood. Distributions slightly differ given different base rate conditions. c) Example where posterior over structures differs among models (assuming a regular base rate). Curved arrows indicate the true underlying generative process unknown to the models.

constraints. It is based on the simulation-and-summary-statistic idea (also written as "summary-statistic" for short), which is as an important approach in Approximate Bayesian Computation in statistics (Blum et al., 2013; Lintusaari et al., 2017; Sunnåker et al., 2013; Y. Zhao et al., 2023). We explore this idea's cognitive plausibility as an explanation for human judgments in our setting. Our model incorporates three features of bounded inference that are often highlighted in cognitive psychology: mental simulation, local computation, and temporally local evidence.

Mental simulation

The tendency to rely on simulation-based approximation to exact inference has been hypothesized to play an important role in model-based reasoning in many scenarios, including physical scene

understanding (Battaglia et al., 2013; Ullman et al., 2018; Hamrick et al., 2016), mechanical reasoning (Hegarty, 2004), and causal judgment (Gerstenberg et al., 2021, 2017). The idea is that instead of computing the likelihood of a potential generative model producing observed data exactly, people instead compare their observations to mental simulations of what *kind of pattern* they expect to happen under different generative models.

Critical to this process is the identification of a useful set of easily tracked abstract cues or features with which to compare such simulations to observations. When a scenario of interest involves complex dynamics, direct surface-level (i.e. "pixel-level") comparison between simulated and observed evidence is generally inappropriate for measuring the likelihood of a hypothesis. Ullman et al. (2018) combined the ideas of simulation and abstraction to model inferences about the latent properties of physical objects (such as masses and forces) from observed dynamics. As a simple example, if imagined heavy objects tend to move more slowly than imagined light objects, this licenses the use of speed as a (fallible) cue to mass.

Concretely, we explore whether simple salient local features of event sequences that are diagnostic (if fallible) guides to local causal relationships can explain human judgments better than a fully Bayesian treatment. The implied cognitive process is that learners draw on (imagined) evidence under different causal ground truth structures in order to develop statistical cues that can be directly applied to pairwise causal judgments. Here we simply investigate two straightforward and salient cues that people might be sensitive to in the current task:

1. **Delay:** The interval between a cause component's activation and the next subsequent effect activation.

2. **Count**: The number of activations of the effect after the cause activation within some time window.

These cues are hand-engineered, and far from exhaustive. However, they are simple to track and turn out to discriminate reasonably well between different types of causal connections. As shown in Figure 5.3b, for the delay cue, we generally expect to see shorter intervals between a control component's activation and the target component's next activation if the control component is a generative cause, a medium and more variable interval if there is no connection or a longer interval if it is preventative. For the count cue, more effect activations are likely to follow the activation of the generative component on average because of the existence of base rate activations as well as generation. In contrast, fewer activations are likely to follow the activation of preventative components. The former cue considers concrete delay information but ignores the possibility of different causal pathways, while the latter cue ignores the exact temporal interval between events (cf. Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018).

(Structurally) local computation

Both the count and delay cues introduced above ignore surrounding structure and context leading to the potential for interference. For example, in the presence of a known preventative cause that has just occurred, an ideal learner should reduce their expectation that a generative cause would produce a short delay to the next event, or a high subsequent effect count. Thus, this approach also captures a principle of local computation (Bramley, Dayan, et al., 2017; Bonawitz et al., 2014; Fernbach & Sloman, 2009; Markant et al., 2016; Griffiths et al., 2015; Davis et al., 2020), predicting that learners will make causal attributions at the level of individual links without accommodating the global context and the full space of global causal models.

The other reason why we apply local computation to this process-level model is that it can greatly reduce the computational cost compared to the global computation approach. In the current continuous-time setting, interventions could happen at any time making every context unique. This means that conditioning one's inference on even a single previous intervention requires learners to simulate a much larger number of one-off context-specific situations. Introducing more of this context sensitivity (i.e. constructing separate summary statistics for each possible combination of causes) would allow a summary-statistic approach to perform closer to normative inference but at the cost of increasing computational demands and reducing generality beyond the set of contexts considered.

(Temporally) local evidence

The final cognitive feature we consider is related to how people parse and segment the evidence encountered across an extended observation of a causal system. Ullman et al. (2018) applied a summary-statistic approach to short observations (5 seconds) and allowed participants unlimited replay opportunities, so assumed people could use cues based on the entire observation. In the scenario considered here, the learner observes causal dynamics for considerably longer (20 seconds, containing dozens of events) without recourse to replays. In general, we experience the world in a single ongoing timeline. Thus, with finite short-term memory storage and attention, it seems plausible that people abstract cues more locally than from full observation. In other studies, people are found to often use temporally local (i.e. recent) information to drive causal model learning (Bramley, Dayan, et al., 2017; Davis et al., 2020; Rehder et al., 2022). Furthermore, people are often unable to recall older evidence exactly (Harman, 1986; Bramley et al., 2015), rather remember whatever conclusions they have drawn on the basis of it.

In line with these ideas, we hypothesize that people segment their observations as they unfold, using recent events to update their beliefs and then discarding their memory of them. We consider two ways to segment continuous-time evidence. As shown in Figure 5.3b, a unit of evidence under both approaches begins with an intervention (i.e. the activation of a control component), capturing the basic principle that causes can only influence what happens later. An *Intervention-window*

segmentation approach treats one unit of observation as the interval between one intervention and the next. This removes the distraction of other interventions that might also influence the effect, but ignores the fact that these interventions might be performed irregularly or reactively, and also that actual effects may not have been revealed before the occurrence of the next intervention. A *Fixed-window* approach ends one unit of observation after a fixed amount of time. This has the advantage of stability in its odds of including all relevant effects' but instead opens the door to confounding influences when subsequent interventions occur within the preceding observation window. A fixed window approach also implies some degree of parallel processing since fixed-length attentional windows may easily overlap in a single timeline.

5.4.4 Summary of modeling frameworks

In sum, we have laid out two approaches to solving the current learning problem. The normative model utilizes the exact timing information of each event, considering all possible observation-consistent ways in which the effects might have been generated or prevented. The summary-statistic model compresses the information by abstracting useful cues and comparing the similarity between cues summarized from observation with mental simulation. We do not see the two accounts as fundamentally in tension. Rather, the summary-statistic approach embodies a set of algorithmically plausible steps to approximate the normative solution.

Given the information compression and the local focus of the summary-statistic approach, its predictions diverge from the normative one in some situations. One example is shown in Figure 5.3c. When B activates and then A activates followed closely by two effects, the normative learner finds this most consistent with the structure where B is a generative cause because the delay between B and the first effect is consistent with its delay expectation, while the other effect could easily be due to the base rate. For the summary-statistic models, the intervention-window approach suffers from a blocking effect, where the occurrence of A masks any potential link between B and the effects. The fixed-window approach suffers from a local computations error, where each effect is potentially attributed to both A and B leading to a marginal preference for the model with both A and B as generative causes. We will show more similarities and differences between the two modeling approaches alongside human behavior in Results sections.

5.5 Overview of experiments

We now report on three experiments that investigate how people infer preventative and generative causal structures in continuous time. Each experiment includes stimuli generated from each of the nine underlying structures we consider (Figure 5.1e). Experiment 1a and 1b aimed at exploring how overall structure and regular and irregular base rates influence causal judgments. Experiment 2 additionally includes stimuli designed to probe whether people make specific mistakes

predicted by the summary-statistics model. All pre-registrations, materials, data, and analysis code are available at https://osf.io/q8n72/. Stimuli for all experiments can be viewed at https://github.com/tianweigong/causal_diamond.

5.6 Experiment 1

5.6.1 Methods

Participants

One hundred and eighty-seven participants from Amazon Mechanical Turk were recruited and reported for Experiment 1a (81 female, 105 male, 1 non-binary, aged 37 ± 11, regular vs. irregular condition: 93 vs. 94) and another 123 participants were recruited and reported for Experiment 1b (45 female, 78 male, aged 39 ± 11, regular vs. irregular: 63 vs. 60). The sample size of Experiment 1a was determined by a power analysis comparing two between-subject groups anticipating a medium sized effect ($d = 0.5$) with a goal of .90 power at the standard .05 alpha. The sample size for Experiment 1b followed a pilot study (Gong & Bramley, 2020) given that both of them aimed to compare participants' performance with normative and heuristic models. Nine additional participants in Experiment 1a were recruited but excluded prior to analysis because they clicked (to respond) more than 300 times during the task (as average participants acted 113±26 times). Hence, we suspected these respondees were either inattentive or non-human. Four additional participants in Experiment 1b were excluded prior to analysis because they clicked more than 300 times during the task (n=2), or failed to pass at least one of two attention questions (n=2).[3] Participants were paid between $1.00 and $2.08 depending on their performance (see below) and experiments lasted around 20 minutes.

Design & Procedure

Overview In both Experiment 1a and 1b, participants judged the causal structure of 18 causal devices (Figure 5.1e). When a generative cause event occurred, it would produce an effect event after 1.5 ± 0.5 s (see Figure 5.2a). Whenever a preventative cause event occurred, any upcoming effect events in the subsequent 3 ± 0.5 s were canceled (see Figure 5.2b). Each base rate event occurred 5±0.5 s after the previous one in the regular base rate condition, or 5±5 s in the irregular base rate condition. The choice of generative delay was based on past studies that suggest people only reliably attribute causal relations to delays of up to around 2 seconds in the absence of context information shaping delay expectations (Shanks et al., 1989; Shanks & Dickinson, 1991).

[3] We also pre-registered to exclude participants who took more than six attempts to pass all instruction comprehension check questions. However, with the benefit of hindsight, we recognized that even attentive participants often required several attempts to pass our stringent comprehension checks. Thus, we opted to relax this exclusion criteria.

We chose the size of the true preventative windows and base rates such that base rates are generally lower than casual influences (i.e. activity is relatively sparse without any generative events) and preventive influences last long enough to have a reasonable chance of preventing something. The true sampled causal delays are unknown to the learner (human or model), but for simplicity we pre-trained (Experiment 1a) or told (Experiment 1b) participants about typical patterns of base rate activations and about typical generative delays and preventative durations in an instruction phase, and so also assumed these parameters were available to all models.

For each device, participants clicked a "Start" button to watch the clip. Each clip started with a base rate activation of the target component and included three pre-set interventions on A and three on B randomly spaced and intermingled over 20 seconds. After that, the clip would end and no further activations could be observed. Components' activations were displayed as the component "lighting up" by changing from gray to yellow for 350 ms. The activation of the control component was accompanied by a hand symbol (Figure 5.1b) and participants were told that this showed that control components were being intervened on by someone or something external to the system, meaning that the interventions happened at random moments rather than following any informative pattern. Clips were selected to make sure that no activation was masked by another on the same component in the clips, and participants were also told about this rule.

Participants were invited to mark their guesses about the two connections during or after the clip by clicking the space between the components (Figure 5.1d). Each clip could only be played once. The order of 18 trials, as well as the click pattern (whether they would have to click once, twice, or tree times to select generative, preventative or non-causal), and the vertical position of A and B components (above or below) were randomized independently between participants.

Participants were informed of the timing of three types of connections as well as the target component's self-activation prior to the inference task. For the base rate specifically, participants in the regular condition were told that the target component would activate regularly about every five seconds and they saw an illustration with a circular arrow to create the impression of periodic activation (Figure 5.1f). Participants in the irregular condition were told that the target component can activate by itself at completely random times and they saw an illustration with an exogenous link intended to imply that someone sometimes activates the target component directly but one cannot anticipate when it will happen (Figure 5.1f). In order to similarly provide timing information, participants were told the base rate activation happens about 2-7 times per clip. Participants had to pass introduction check questions before starting the experiment. To properly incentivize accurate judgments, a 3-cent bonus was paid for each correctly identified connection and non-connection during the main task in addition to the basic $1 payment.

Experiment 1a In Experiment 1a, to generate stimuli from different structures (e.g. both generative, one generative and one non-causal) and different conditions (i.e. regular vs. irregular) comparable, we used a Latin-square design. We first created 18 causal delay seeds independently.

Each of these included a set of timings for interventions, base rate activations, which depended also on whether the base rate was regular or irregular, and what generative delays (or blocking windows) A and B would have if they were generative (or preventative) components. Under each seed, 18 stimuli (9 causal structures $\times$ 2 base rate settings) were generated by implementing generative or preventative influences according to the ground-truth structure (see Figure 5.4a for an example of a single seed manifesting under each structure and base rate condition). Across different seeds, the timing and order of interventions were randomly generated to capture the diversity of ways in which the interventions could be interleaved ranging from perfectly interleaved (e.g. $ABABAB$) to perfectly clustered (e.g. $AAABBB$, see Figure 5.4b for an example of a single structure under different seeds and base rate conditions). All stimuli were finally divided into 18 sets (9 sets for each base rate setting) according to a Latin-square design that ensured participants would only see one structure under each seed (see https://osf.io/sqv6c for the counterbalancing matrix). Participants were randomly assigned to one of these 18 sets.

In the instructions, participants saw training videos that showed the patterns of the target component's base rate activations (corresponding to their condition) and also what happens after intervening on a causal system with a single (generative, non-causal, or preventative) connection. They completed a single practice trial in which the true causal device included one generative connection and one non-causal connection. Feedback was provided in the practice trial but not in the test trials.

Experiment 1b Experiment 1b differed from Experiment 1a from two perspectives. Firstly, although we assume the provenance of the summary-statistic approximation to be mental simulation, cues might also be derived from experience with the "labeled data" included in the instructions or practice trials. Therefore, Experiment 1b only kept the text instructions and removed the training videos and practice trials, to show that labeled data were not necessary for participants to complete the task.

Additionally, given that the stimuli in Experiment 1a were generated by one of the ground truth structures, the normative model and summary-statistic approximations often made similar predictions. To probe how participants react to situations with stronger discrepancies between the normative and summary-statistic predictions, we created some stimuli that were not generated by any particular causal device. We created two blocks of stimuli in Experiment 1b. Block 1 included nine stimuli for each participant, which replicated the procedure of Experiment 1a, and served to ensure participants were habituated to reacting to "normal" stimuli. In Block 2, we generated potential test stimuli by randomly distributing six interventions and between 1 and 9 effects across a 20 second trial. We selected sequences for which the structure predictions of the normative and summary-statistic models were strongly dissimilar, while ensuring that these stimuli were not too normatively improbable (i.e. that they could conceivably have been generated by one of the

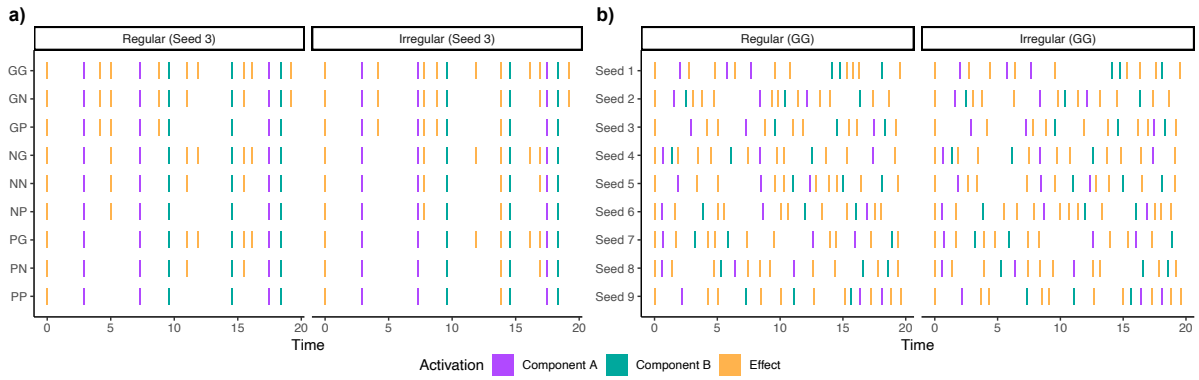


Figure 5.4: a) Examples of a single seed under different structures and base rate conditions (from one stimulus seed used in Experiment 1a). Y-axis refers to the roles of Component A and B (e.g. GP: A is a generative cause and B is a preventative cause). b) Examples of a single structure manifesting under different seeds and base rate conditions.

causal structures). [4] There were 27 stimuli for each condition and each participant observed nine of them. Block 1 always preceded Block 2 so that the first half task would be identical to Experiment 1a. Participants completed 18 trials in sequence without any delineation between the blocks. All other experimental settings remained identical to Experiment 2. The bonuses were, in reality, determined by doubling the bonuses participants gained in Block 1.

5.6.2 Results

We focus on analyzing participants' accuracy by comparing their judgments against the ground truth. We investigate whether participants' performance was influenced by the nature of the underlying causal mechanism, base rate regularity, or the observed intervention sequence (i.e. whether this involves interleaved interventions on the two components or clusters of interventions on one component then the other). Since these analyses require there to be a correct answer, for Experiment 1b we only include Block 1.

To compare our models' behavior qualitatively with participants', we simulate judgments of each model type after observing the same stimuli as the participants. We used a fitted softmax parameter for each model and repeated each simulation 200 times per participant to obtain stable and consistent distributions of simulated judgments (see Appendix A.3 for model fitting details). For summary-statistic models, we average predictions under the two proposed features with equal weights to form a combined prediction (cf. Ullman et al., 2018). Results of intervention-window vs.

[4]Specifically, we picked the stimuli where at least one (intervention-window) summary-statistic cue (Delay or Count) had a different dominant answer compared to the normative model and rejected any for which the likelihood of the most probable structure producing the data was extremely low ($< 10^{-40}$). The squared error between normative and summary-statistic predictions in Block 1 (trials with the ground truth) and Block 2 (trials without the ground truth) was 0.22 vs. 0.53 on average. The likelihood of the most probable structure according to the normative model in Block 1 and Block 2 was 0.07 vs. 0.004 on average.

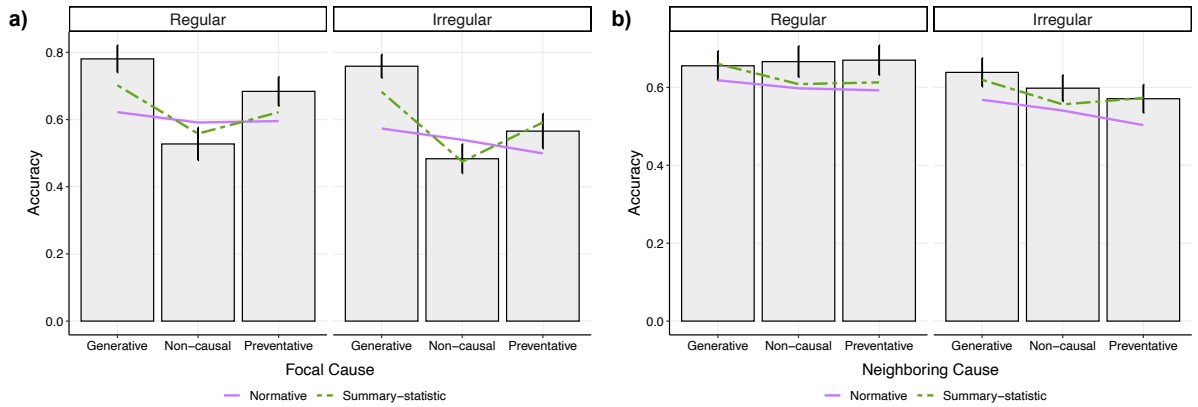


Figure 5.5: a) Accuracy of different causal connections in Experiment 1. b) Accuracy in judging a connection (averaged across generative, preventative, or non-causal target connections) when paired with different types of connections in Experiment 1. Lines indicate the performance of simulated normative and summary-statistic learners each with a fitted softmax parameter based on all participants' data in Experiment 1 (see Appendix A.3). Error bars indicate 95% confidence intervals.

fixed-window summary-statistics were similar at the aggregated level, and hence we only visualize the intervention-window results in the figures.

Overall performance

In Experiment 1a, participants in both regular and irregular conditions performed above chance at the connection level (chance = 33%, regular: 66%±22%, $t(92) = 14.75$, $p < .001$, $d = 1.53$, 95%CI of $d = [1.23, 1.84]$; irregular: 61%±18%, $t(93) = 14.67$, $p < .001$, $d = 1.52$, 95%CI of $d = [1.22, 1.82]$) as well as the structure level (1 = correct in both connections; 0 = otherwise, chance = 11%, regular: 49%±27%, $t(92) = 13.83$, $p < .001$, $d = 1.43$, 95%CI of $d = [1.15, 1.73]$; irregular: 41%±22%, $t(93) = 13.27$, $p < .001$, $d = 1.37$, 95%CI of $d = [1.09, 1.66]$). These patterns were replicated in Experiment 1b, where participants also performed above chance at both connection (regular: 67%±22%, $t(62) = 11.93$, $p < .001$, $d = 1.50$, 95%CI of $d = [1.15, 1.88]$; irregular: 59%±19%, $t(59) = 10.11$, $p < .001$, $d = 1.30$, 95%CI of $d = [0.96, 1.66]$) and structure levels (regular: 49%±29%, $t(62) = 10.64$, $p < .001$, $d = 1.34$, 95%CI of $d = [1.00, 1.69]$; irregular: 39%±23%, $t(59) = 9.21$, $p < .001$, $d = 1.19$, 95%CI of $d = [0.86, 1.53]$). Indeed, accuracy did not differ between Experiment 1a and 1b at the connection level (regular: $t(154) = 0.09$, $p = .926$; irregular: $t(152) = 0.81$, $p = .418$) or the structure level (regular: $t(154) = 0.004$, $p = .997$; irregular: $t(152) = 0.56$, $p = .578$). This means that labeled data in the form of video training and practice trials were not a necessary condition for participants' success in this task. We therefore combine stimuli from two experiments in later analyses to obtain a larger sample size.

Focal and neighboring causes

To investigate participants' ability to identify generative, non-causal, and preventative connections, as well as whether the base rate regularity or the neighboring connections would influence performance, we performed a 3 (focal cause: generative, non-causal, preventative) × 3 (neighboring cause: generative, non-causal, preventative) × 2 (base rate regularity: regular, irregular) mixed ANOVA. Each trial provided two data points here, one regarding A as the focal cause and B as the neighboring cause and the other regarding B as the focal cause and A as the neighboring cause.

There was a main effect of focal cause ($F(2, 616) = 101.24$, $p < .001$, $\eta_p^2 = .247$, 95%CI of $\eta_p^2 = [.200, .293]$). Participants performed best at identifying generative connections (77%±24%), then preventative connections (63%±31%), and finally non-causal connections (51%±29%, Figure 5.5a). The differences were all pairwise-significant (Bonferroni adjusted $p < .001$). [5]

There was a main effect of base rate regularity ($F(1, 308) = 7.07$, $p = .008$, $\eta_p^2 = .022$, 95%CI of $\eta_p^2 = [.003, .057]$). Participants tended to perform better in the regular ($66\% \pm 22\%$) than the irregular ($60\% \pm 19\%$) condition. However, there was an interaction between focal cause and base rate regularity ($F(2, 616) = 3.69$, $p = .026$, $\eta_p^2 = 0.012$, 95%CI of $\eta_p^2 = [.001, .028]$). Analysis of the simple effects showed that the regularity difference was only significant for preventative causes (Figure 5.5a). This is consistent with the principle that identifying preventative causes relies heavily on having a good counterfactual expectation of what would have happened in the causal system in the absence of the focal cause.

The main effect of neighboring cause was non-significant ($F(2, 616) = 2.76$, $p = .064$) while there was a interaction between neighboring cause and base rate regularity ($F(2, 616) = 6.66$, $p = .001$, $\eta_p^2 = .021$, 95%CI of $\eta_p^2 = [.005, .042]$). The neighboring connections made a difference in the irregular condition ($F(2, 308) = 8.56$, $p < .001$, $\eta_p^2 = .053$, 95%CI of $\eta_p^2 = [.017, .095]$), but not in the regular condition ($F(2, 308) = 0.41$, $p = .662$, Figure 5.5b). Participants in the irregular condition performed better when the neighboring connection was generative than non-causal ($t(308) = 2.48$, $p = .041$, $d = 0.099$, 95%CI of $d = [0.003, 0.195]$) or preventative ($t(308) = 4.13$, $p < .001$, $d = 0.165$, 95%CI of $d = [0.069, 0.261]$). This means that when the base rate was uncertain, a generative cause could stand in by setting up strong expectations. Other two-way or three-way interactions were non-significant ($ps > .05$).

For simulated model-based learners, the summary-statistic learner exhibited a similar tendency as participants, performing worse in identifying non-causal connections (Figure 5.5a). The accuracy of both normative and summary-statistic learners was partly dependent on the neighboring cause. As shown in Figure 5.3b, the summary-statistic distributions of the non-causal type,

[5] To rule out that this main effect was merely due to people generally selecting more answers as generative and preventative than non-causal, we calculated the F1-score for each cause (Powers, 2011). The patterns were the same when using the F1-score as the index ($F(2, 540) = 181.89$, $p < .001$, $\eta_p^2 = .403$, 95%CI of $\eta_p^2 = [.352, .448]$).

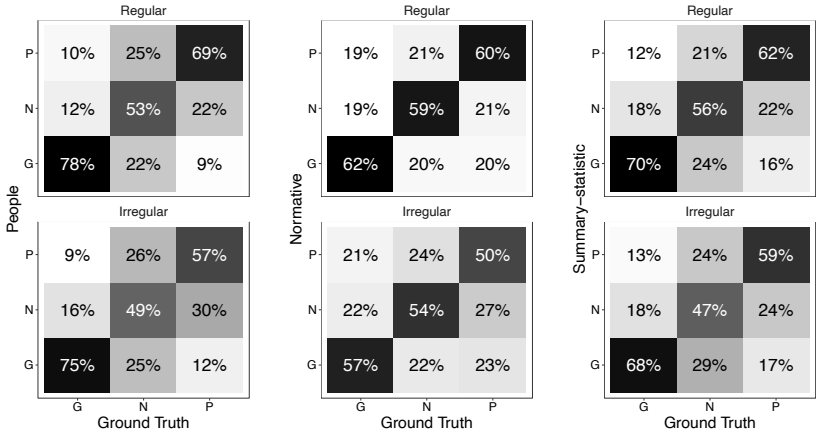


Figure 5.6: Confusion matrices for participants' and models' choices under different ground truths in Experiment 1. The normative and summary-statistic learners were simulated with a fitted softmax parameter based on participants' data in Experiment 1.

particularly the Delay distributions, frequently exhibit overlaps with both other distributions, and furthermore, the other types (generative or preventative) typically have higher density in the overlapping region.

Confusion matrices

Figure 5.6 shows the proportion of participants' choices under different ground truths. We explored the frequency of choices when people made inconsistent judgments with the ground truth. Under the regular base rate, people were equally likely to judge a generative connection as a non-causal one or a preventative one (12% vs. 10%, chi-square goodness of fit: $\chi^2(1) = 3.01$, $p = .082$). They were equally likely to judge a non-causal connection as a generative or preventative one (22% vs. 25%, $\chi^2(1) = 2.65$, $p = .103$) while they more often judged a preventative connection as a non-causal one than a generative one (22% vs. 9%, $\chi^2(1) = 83.41$, $p < .001$). The results of irregular base rate were similar (non-causal ground truth: 25% vs. 26%, $\chi^2(1) = 0.70$, $p = .404$; preventative ground truth: 30% vs. 12%, $\chi^2(1) = 107.96$, $p < .001$) except now participants also more often judged a generative connection as non-causal than preventative (18% vs. 9%, $\chi^2(1) = 29.55$, $p < .001$). The summary-statistic learner exhibited a similar tendency to human participants, tending to mistake preventative or generative connections more often as non-causal, rather than mistaking one for the other.

Intervention order

We examined the influence of the intervention sequence. The intervention patterns in the experimental stimuli were randomly generated (albeit balanced to include 3 interventions per control component) and hence varied in terms of the sequence. In some trials, participants observed data

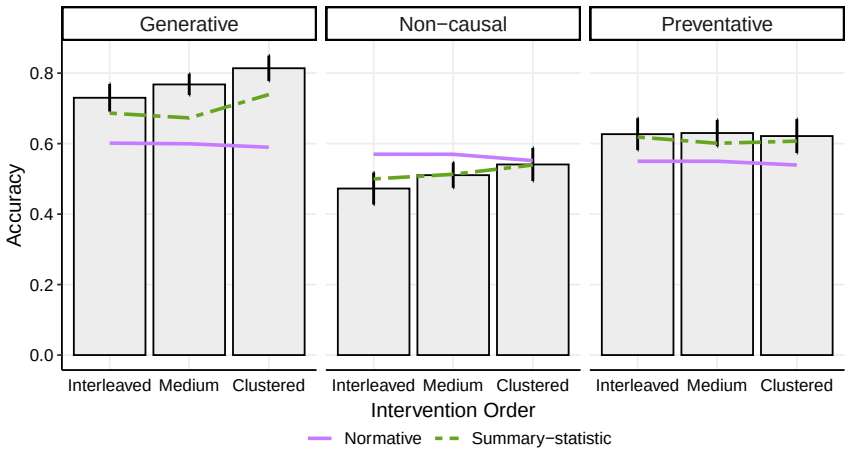


Figure 5.7: Accuracy separated by intervention order in Experiment 1. Lines indicate the performance of simulated normative and summary-statistic learners each with a fitted softmax parameter based on participants' data in Experiment 1. Error bars indicate 95% confidence intervals.

in which interventions on one *component* were "interleaved" (e.g. A in $ABABAB$ or $ABBABA$), in others they were fully "clustered" (e.g. A in $AAABBB$ or $BAAABB$), and in others they were partially clustered (e.g. A in $AABABB$ or $ABBBAA$) which we called a "medium" level. We performed a 3 (focal cause: generative, non-causal, preventative) $\times$ 3 (intervention order: interleaved, medium, clustered) $\times$ 2 (base rate regularity: regular, irregular) mixed ANOVA. Each trial provided two data points, one regarding A as the focal cause and the other regarding B as the focal cause. The effects regarding focal cause and base rate regularity were similar to previous analyses and hence we only focus on the effects related to intervention order here.

There was a main effect of intervention order ($F(2, 538) = 9.39$, $p < .001$, $eta_p^2 = .034$, 95%CI of $\eta_p^2 = [.012, .061]$) and an interaction effect between intervention order and focal cause ($F(4, 1076) = 3.22$, $p = .012$, $\eta_p^2 = .012$, 95%CI of $\eta_p^2 = [.001, .022]$). As shown in Figure 5.7, the clustering intervention mainly benefited the identification of generative ($F(2, 269) = 11.40$, $p < .001$, $\eta_p^2 = .078$, 95%CI of $\eta_p^2 = [.031, .131]$) and non-causal ($F(2, 269) = 3.76$, $p = .025$, $\eta_p^2 = .027$, 95%CI of $\eta_p^2 = [.002, .063]$) connections, while the effect was insignificant for preventative connections ($F(2, 269) = 0.07$, $p = .935$). The summary-statistic learner demonstrated a similar influence of the intervention order as humans, while the normative learner performed indifferently across different intervention orders (Figure 5.7).

Trials optimized for model discrimination

Block 2 in Experiment 1b contained stimuli that were not generated from a particular ground truth structures, but rather generated so as to distinguish strongly between normative and summary-statistic models. Figure 5.8 shows the choice proportion of human learners vs. simulated learners on each stimulus. The choices simulated from the summary-statistic model were better correlated

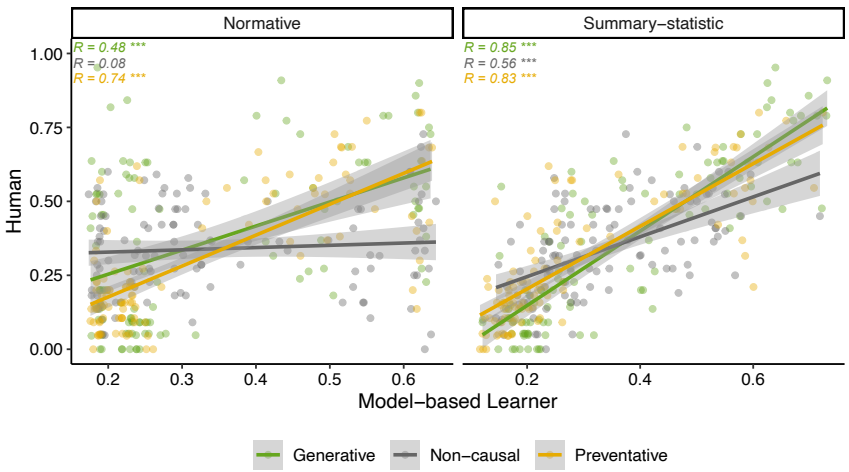


Figure 5.8: Scatterplots of simulated model-based learners predictions and human judgments on the proportion of choosing different causal types in stimuli with no ground truth in Experiment 1b. Each connection in a stimulus is represented by three data points in the figure corresponding to the participant's and models' average probability assigned to that possibility. The normative and summary-statistic learners were simulated with a fitted softmax parameter based on participants' data in Experiment 1. Error bars indicate 95% confidence intervals.

with human judgments across generative, non-causal, and preventative answers. In particular, the summary-statistic model captured when people tended to judge a variable as non-causal (gray points and line) which often diverged from the normative prediction.

Model fitting

To check quantitatively how well the models we have considered capture participants' causal judgments, we fit all participant judgments with our normative and summary-statistic models at both aggregate and individual levels. The details of the model fitting procedure can be found in Appendix A.3.

Participants choices were best captured by the summary-statistic approach, specifically by the variant that segments evidence according to the intervals between interventions (Table 5.1). This is corroborated by the individual level fits, where the largest proportion of participants were fit by *summary-statistic (intervention based)* in both regular and irregular conditions across experiments (model fits separated by conditions are shown in Table A.1).

We provide additional model fitting results in Appendix A.4. In Table A.2 we fit answers from Experiment 1b separated by blocks. The difference in cross-validation log-likelihood or BIC between normative and summary-statistic models was more pronounced in the no-ground-truth block than in the ground-truth block, which reflected that people's judgments were indeed more similar to the summary-statistic model. In Table A.3 we fit participants' answers with each cue separately to see whether they were dominated by Delay or Count rather than their combination. Results indicate that models with one or another cue did not fit participants' judgments better

Table 5.1: Model fits.

	CV	BIC	τ	N (Regular)	N (Irregular)
Experiment 1a					
Normative	-6054	12110	0.44	17%(14%)	19%(13%)
SS (intervention-window)	**-5857**	**11718**	0.23	**53%(47%)**	**46%(45%)**
SS (fixed-window)	-5998	12002	0.30	19%(17%)	26%(21%)
Random	-7430	14859		11%(22%)	10%(21%)
Experiment 1b					
Normative	-4426	8833	0.58	14%(11%)	10%(7%)
SS (intervention-window)	**-4054**	**8113**	0.23	**60%(51%)**	**58%(52%)**
SS (fixed-window)	-4167	8338	0.33	21%(17%)	25%(25%)
Random	-4887	9774		5%(21%)	7%(17%)
Experiment 2					
Normative	-1058	2113	4.43	2%(0%)	
SS (intervention-window)	**-948**	**1835**	0.19	**50%(50%)**	
SS (fixed-window)	-955	1915	0.34	43%(40%)	
Random	-1059	2119		5%(10%)	

Note: SS refers to summary-statistic models. The "N (Regular)" and "N (Irregular)" columns display the proportion of individuals best-fit by each model according to CV, with BIC results in the brackets.

than models that mixed two cues. In Figure A.1, we performed a grid search in $[1, 7]$ seconds with a step of 0.5 s to test whether the fixed-window model fits were sensitive to our choice of a 4 second window. Models with different fixed-window lengths always had substantially larger BICs than the model with the inter-intervention window approach, meaning that, even had we fit window length as an additional parameter it would not outperform by-intervention segmentation in describing participants. This was true despite the fact that the models' accuracy in causal identification is quite sensitive to the window length.

5.6.3 Discussion

In Experiment 1, we showed that people are capable of using temporal information to learn causal structures that involve generative and preventative relationships. It also showcases several interesting differences between generative and preventative causation, which we return to in the General Discussion. Human judgments were better aligned with the summary-statistic models' predictions in both quantitative results and aggregate qualitative results. Nevertheless, the data in Experiment 1 was complicated, meaning we can do more to distill simpler, more intelligible examples of how the normative and summary-statistic models diverge in their judgments. In Experiment 2, we examine judgments about minimal event sequences for which the summary-statistic and normative learners differ in their dominant answers.

5.7 Experiment 2

We designed two types of stimuli for which two models have different dominant answers. They are based on the two locality principles driving the summary-statistic model: (1) Local computation; meaning summary statistic learners fail to account for the influence of the other connections in the system, and (2) Local evidence; meaning summary statistic learners fail to take into account whatever happened before their current observation window. For the first type of stimuli we use scenarios where a learner needs to identify a generative target cause that is paired with a preventative cause. This presents a challenge for local computation because the preventative cause can block the generative causes' influence and mislead a local learner into believing the target connection is a non-causal connection, because it is statistically associated with fewer events per window or longer delays than generative causes have on average across the task. The second case type is scenarios where a non-causal target is paired with a generative neighboring component. For a local learner who only focuses on a small time window after each intervention, the generative influences can easily spill over to the observation window during which the learner is focused on the target non-causal component and leading to statistics more typical of generative causation, because it is associated with more events and shorter delays than non-causal components exhibit on average across the task. Experiment 2 focused on the regular base rate condition, since this yields the larger predicted difference between normative and summary-statistic based judgments, though we also checked that the dominant answer for each model was the same under the irregular base rate.

5.7.1 Methods

Participants

60 participants from Prolific were recruited and reported (32 female, 28 male, aged 41 ± 12). The sample size was determined by a power analysis assuming a medium sized effect ($d = 0.5$) in comparing within-subject judgments on the target cause and the goal of .90 power at the standard .05 alpha. No participants were excluded from this experiment based on the criteria we pre-registered.

Design & Procedure

Participants' task was very similar to the regular condition in Experiment 1b, where they needed to judge the roles of two connections given a 20-second clip of evidence. No video training or feedback was provided. The hand-crafted stimuli are shown in Figure 5.9. For each stimulus, we call one component the "target", and the other the "lure", which could affect participants' judgments about the target. Each clip contained two segments of evidence where the two components activated close together, so their influences on the system (if any) were misleading to the summary statistic

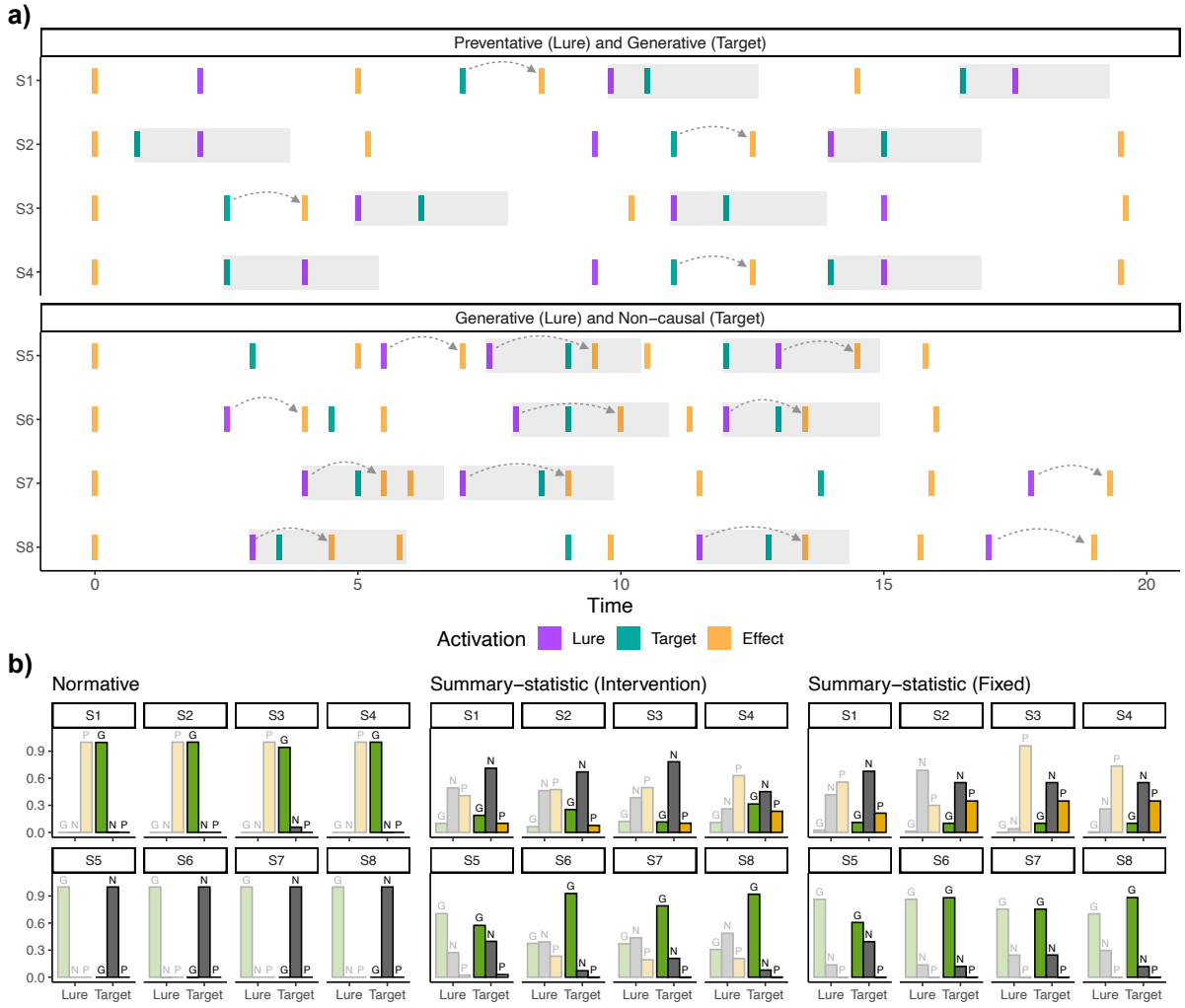


Figure 5.9: Stimuli and model predictions in Experiment 2. a) Stimuli. Curved arrows indicate the true underlying generative process. b) Judgment predictions from different models. The normative and summary-statistic models particularly differ in their judgments about the target components, with opaque bars used to highlight where the modal response shifts between normative and summary statistic models.

model (gray shadows in Figure 5.9), but also contained evidence where the components occurred far enough apart to make the true structure recoverable by the normative model.

We constructed four exemplars of the two stimulus types (Figure 5.9). For the *PG* type (preventative lure and generative target), the lure often cancels the influence of the target, and hence the summary statistics of the target are more aligned with the non-causal summary statistics. For the *GN* type (generative lure and non-causal target), the lure's influence spills over into the observation window of the target, leading to summary statistics more consistent with a generative target component. Therefore, the summary-statistic approach predicts systematic errors in these cases that are not predicted by the normative model (Figure 5.9).

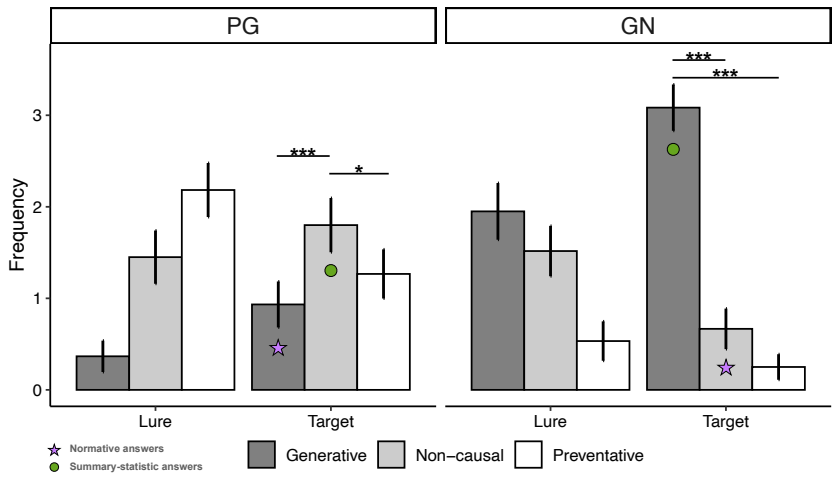


Figure 5.10: Judgments of two types of stimuli in Experiment 2. Each type included four stimuli. Participants' dominant answers for the target component are consistent with the dominant answers from the summary-statistic model (the green dots) rather than the dominant answers from the normative model (the purple stars). Error bars indicate 95% confidence intervals.

Participants went through 6 practice trials sampled from Experiment 1 (with structures GG, NG, GP, NN, PN, PP) before 8 testing trials, to ensure that they had some experience with different structures and edge types under more normal conditions. The vertical positions of two control components (above or below) were randomized across trials. The order of trials was randomized within the practice and testing phases. Participants completed 14 trials in sequence without any delineation between the practice and critical trials. The bonuses were, in reality, administered proportional to the bonuses participants gained in the practice phase (given that we predicted participants would make systematic errors in the test phase).

5.7.2 Results & Discussion

For the PG stimuli, participants judged the targets as non-causal 1.8±1.1 times on average out of 4 trials (above the 33% change level, $t(59) = 3.15$, $p = .003$, $d = 0.41$, 95%CI of $d = [0.14, 0.67]$). More importantly, participants judged them more often as non-causal than generative ($t(59) = 3.62$, $p < .001$, $d = 0.82$, 95%CI of $d = [0.44, 1.19]$) or preventative ($t(59) = 2.11$, $p = .04$, $d = 0.49$, 95%CI of $d = [0.12, 0.85]$). For the GN stimuli, participants judged the targets as the generative one 3.1±1.0 times on average out of 4 trials (above the 33% change level, $t(59) = 6.03$, $p < .001$, $d = 0.78$, 95%CI of $d = [0.49, 1.07]$). Meanwhile, participants judged them more often as generative than non-causal ($t(59) = 10.64$, $p < .001$, $d = 2.63$, 95%CI of $d = [2.13, 3.11]$) or preventative ($t(59) = 16.50$, $p < .001$, $d = 3.58$, 95%CI of $d = [3.00, 4.16]$). This means that for both kinds of stimuli, participants' dominant answers lined up with the summary-statistic models and diverged from those of the normative model.

The model fitting results are shown Table 5.1. Similar to Experiment 1, participants' answers were better fit by the summary-statistic models than the normative model. In general, they were also better aligned with the intervention-window segmentation than the fixed-window segmentation. This is also supported by a qualitative result that for GN stimuli, both the intervention-window model (Figure 5.9) and participants (Figure 5.10) regarded the lure as less likely to be a generative cause than the target component ($t(59) = 5.56$, $p < .001$, $d = 1.04$, 95%CI of $d = [0.65, 1.41]$), while the fixed-window model regarded the probabilities as more even (Figure 5.9). When it comes to the individual difference, participants split more evenly across the intervention-window and fixed-window models than Experiment 1, which may imply that some participants do consider longer windows in situations when interventions interleaved heavily and hence evidence of intervention-based windows was sometimes too short to rely on.

5.8 General discussion

This paper examined how people infer causal structure on the basis of observing events in continuous time. The project was motivated by the fact that classical causal structure induction research has largely focused on inferences from atemporal statistical information, essentially sidestepping the role of event timing and delay, or else reducing it to a simple sequence of equally spaced measurements. Meanwhile, empirical research (not to mention common sense) suggests people rely strongly on event timing for causal reasoning, using temporal information to guide causal attributions even when it is inappropriate to do so. It seems likely, therefore, that time is integral to our representation of causality and hence deserves careful formal and empirical treatment.

While the space of causal structures we explored was relatively restricted, our task was challenging due to the spontaneous activations of the effect component and potential interactions between generative and preventative cause components. There were always multiple competing explanations for any effect occurrence or surprising non-occurrence, and as such, normative reasoning about the structure behind the evidence required entertaining and marginalizing over many hypothetical mappings between events. Nevertheless, participants' were able to correctly identify the majority of causal components well above chance even when base rate activations of the effect were unpredictable (Experiment 1a and 1b) and even without pretraining about the true causal delays (Experiment 1b). Our experiments thus provide an initial empirical demonstration that people can use real-time temporal information to detangle the influences of generative and preventative causes and identify causal structures involving combinations thereof.

5.8.1 Empirical findings

By including both preventative and generative relationships in our task, we have empirical results showing how the identification of these two types of relationships differ from each other in a continuous-time setting.

First, base rate regularity has a larger impact on identifying preventative relationships than generative relationships. Participants can better identify preventative connections when the effect otherwise activates regularly. This is aligned with the principle that detecting preventative causation relies heavily on one's expectation of what would otherwise have happened in the causal system (Cheng, 1997; Buehner et al., 2003; Griffiths & Tenenbaum, 2005).

Second, when judging a causal connection in the system, the type of neighboring connections matters. Experiment 1 showed that when the base rate is irregular, participants could better identify a connection when it was paired with a generative neighbor rather than a non-causal or preventative neighbor. This can be explained by the fact that a generative connection can increase the predictability of the effect, which is helpful in general but particularly when the base rate is unpredictable. Experiment 2 showed that a preventative neighbor can cancel out a generative influence and mislead people to judge a generative connection as non-causal.

Third, the timing and sequence of interventions matter when making causal judgments, and it affects the identification of generative and preventative connections in different ways. Participants identified generative and non-causal relationships better when the interventions were clustered, rather than interleaved. This makes sense given that the evidence under clustered interventions involves less interference from neighboring connections. We confirmed this in Experiment 2 where we show that deliberately interleaved evidence leads participants to systematically mistake the roles of generative and non-causal connections. In contrast, the advantage of clustered interventions disappeared when it came to prevention. To identify preventative relationships, it makes sense to spread out interventions so their influence covers more of the timeline, and in particular to perform them ahead of whenever one has a strong expectation of the effect occurring (Melchers et al., 2006; Lovibond & Lee, 2021). To our knowledge, these findings represent the first systematic investigation of how human causal judgments engage with a setting where generative and preventative causal influences intertwine and interact in time.

5.8.2 Normative vs. summary-statistics

To better understand how participants made their judgments, we contrasted two learning models: An exhaustive normative account and a summary–statistic-based local approximation. Both accounts were able to identify generative and preventative influences well in our task, but only the summary statistic account could capture cases in which participants were worse at identifying the non-causal connections (Experiment 1) and misled by interleaved interventions (Experiment 1 and

2). Quantitatively, the summary-statistic account also fits participants' judgments across both experiments better.

Our normative model demonstrates that near perfect inversion of the generative causal model is possible for a learner with exactly the correct delay assumptions and unlimited processing power. It works via reasoning at the token level of actual attribution (Halpern, 2016), suggesting this kind of reasoning is key for achieving benchmark performance in this small data setting. The summary-statistics account takes a different approach that is computationally much more frugal and scalable to more complex causal models, but has the cost of being less sensitive to precise event timing, and being more susceptible to interference between components. The approach combines several core principles of bounded cognitive processing: Use of simulation from generative mental models and comparison via summary statistics in place of an exact or intractable likelihood calculation (Battaglia et al., 2013; Ullman et al., 2018; Blum et al., 2013; Lintusaari et al., 2017; Sunnåker et al., 2013). It combines this with local (Fernbach & Sloman, 2009; Bramley, Dayan, et al., 2017; Davis et al., 2020) and incremental (Bramley, Dayan, et al., 2017; Davis et al., 2020; Rehder et al., 2022) processing to break up the global inference problem into a series of spatially and temporally local subproblems. The departures from the ideal of the global normative thinker allow it to explain several error patterns exhibited by participants. In general the normative model serves to showcase the rapidly compounding challenge of maintaining a global perspective when processing evidence that includes multiple causal influences that intertwine and interact in real time (Gong et al., 2023; Bramley, Mayrhofer, et al., 2017).

Imagined experiences are a core feature of our conscious experience and as such, mental simulation has been implicated by a number of theories of cognition as playing key roles in both model-based inference and planning (Battaglia et al., 2013; Hamrick et al., 2016; Ullman et al., 2018; Ludwin-Peery et al., 2020; Gerstenberg et al., 2021; Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018). Mental simulation is thought to be key to offline (Hinton et al., 1995), and simulation phases are now a common part of the training regimen for large Reinforcement Learning Models (Mnih et al., 2015; Ellis et al., 2020). Our experiments add one small piece to this research line, showing how an inference mechanism grounded in simulation and the extraction of summary statistics may explain how people mitigate the computational costs involved in reverse engineering the causal mechanisms that explain the events we observe in real time.

The idea of combining sampling from a generative model with summary statistics stems from Approximate Bayesian Computation (Blum et al., 2013; Lintusaari et al., 2017; Sunnåker et al., 2013). The approach makes it possible to approximate an intractable Bayesian inference by using the similarity between data simulated from a hypothesized model or parameter setting and observed data as a proxy for the likelihood of that model or parameter setting. Choosing the best summary statistics or loss function for a domain is a research area in itself in machine learning (Csilléry et al., 2010), while identifying what summary statistics might be used in cognition is another challenging and unsolved problem. We do not solve this problem here, but simply hand

selected two basic summary statistics (cf. Ullman et al., 2018) on the grounds that they reflect the most basic and easily reported timing measurements people can make in online settings. We showed that the delay and count cues were reasonably diagnostic in our task (Experiment 1) but also unpacked the circumstances under which they can be misleading (Experiment 2).

Within the summary-statistic framework, we considered two ways participants might segment the trials into counting windows. We proposed they might either track events within fixed-length windows after each intervention or use the gaps between each intervention directly as a count window. The inter-intervention segmentation variant captured participants' behavior better despite the fact that the windows were of markedly different lengths detracting from the reliability of the metric. A potential explanation for this is that people may be fundamentally unable to track events from multiple causal perspectives in parallel, thus being forced to rely on the uneven inter-event windows (Davis et al., 2020; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018). Of course, in an active learning context, the learner is free to perform interventions at their own pace. This research suggests that what learners are able to attend to and measure is likely to shape their approach to interventions in time. For instance, one way to make inter-intervention count statistics as powerful as possible is to intervene on a regular schedule, eliminating the confound of episode length, while leaving as large as possible gaps between interventions additionally minimizes spillover effects. Interestingly, these are cognitive rather than normative considerations since the ideal observer is practically indifferent to the regularity of the intervention spacing.

5.8.3 Alternative accounts

One popular recent idea in the causal cognition literature is that people form and adjust causal theories locally and incrementally (Bramley, Dayan, et al., 2017; Bramley, Mayrhofer, et al., 2017; Davis et al., 2020; Markant et al., 2016; Fernbach & Sloman, 2009). For instance, Bramley, Dayan, et al. (2017) model causal structure learning (in discrete trial contexts) as a process of incremental adaptation of a single global hypothesis driven by the need to accommodate new evidence as it arrives. They argue that causal learners do focus locally when grappling with complex structures, but that many are able to condition on their current beliefs about neighboring connections rather than ignoring them altogether, leading to patterns of sequential local focus and anchoring that still tend to favor the correct global structure in the limit. We did not collect the interim judgments we would need to probe this account directly, but we think it is entirely plausible that people focused on the roles of the components not just separately but also serially, perhaps flipping their attention back and forth several times throughout a trial. For example, if participants focused on a generative component first and a preventative component second, they might have been able to take advantage of their expectation of events produced by the apparently generative component to supercharge their inferences about prevention.

The other idea is based on the "smart initialization and short search" algorithm in Ullman et al. (2018). Analogous to our findings, they showed that although human physical learning was better captured by a summary-statistic account than a noisily normative Bayesian model, responses could be even better fit by a mechanism that combines the two. Their best-fit model used the prediction of a summary-statistic approach as a starting hypothesis, and then made local adjustments to this by running a short Markov Chain Monte Carlo search chain. Such a smart initialization could play an important role here too. It is plausible that some participants may have performed similar steps, i.e. forming an impression of the role of a component due to the delays and counts but adjusting this when accommodating a belief about the neighboring connection or an understanding of the regularity of the base rate.

5.8.4 Future directions

To date, causal learning in continuous time has received little attention, meaning there are numerous basic research questions still to be addressed. In the current paper, we focus on just one of these, providing a close examination of the interplay between inference about generative and preventative causal relationships. However, for this we make specific assumptions about the scope with which preventative influences work. Concretely, we conceive of preventative influences as eliminating all expected effects for a short time no matter their cause. However, there are several alternatives that seem at least as salient and may be more appropriate depending on knowledge of the context and mechanisms involved. For example, prevention could work by blocking the next one event (or perhaps the next N events) rather than blocking everything for a fixed window. Prevention could also operate on "links" rather than "nodes" within the causal graph, for example blocking the action of a generative cause on an effect, but leaving the spontaneous activations of that effect intact, or visa versa (Fraser & Holland, 2019; Chow et al., 2023; C. D. Carroll & Cheng, 2009).

In the current learning task, causal influences were represented as operating between point events. This is a major simplification from many real scenarios in which variables involved in causal interactions are often able to take multiple, or even a continuum of, values. The cat in our motivating example might drink more or less water or hold different teaser toys in higher or lower regard leading to faster, slower, more or less intense effects. Even though events are abstractions of continuous inputs, and many, such as state changes, are readily thought of as punctate, many everyday event concepts clearly have non-zero duration and often have internal structure such as a gradual or sudden onset or offset. For example, given enough time, many of the states referred to in causal learning scenarios are not permanent. "Wet ground" dries. "Tanned skin" fades. Many disequilibria will either dissipate or recover without external intervention. Other states, such as a turned-on light bulb may tend to persist until cancelled, i.e. by switching the switch a second time. These could be seen as events with an infinitely long duration (i.e. permanent

state-changes). As event duration reduces, it becomes less likely that events will overshadow one another. Point events are a limiting-case abstraction of this where the duration is reduced to zero, resulting in a setting where there is no true causal overshadowing (Paul & Hall, 2013). That is, generative cause will always produce an observable effect even if it occurs close to another event. However, in settings with longer events it becomes increasingly important to consider the super-secession situations and perhaps to apply the noisy-or or noisy-and-not frameworks (Cheng, 1997; Griffiths & Tenenbaum, 2009) that capture how in contingency settings, effects can easily be hidden due to an already-occurring, or already-prevented target. Future research could study how people represent the duration of causal events as well as their influences and thus begin to form a richer theory of causal concepts in time that captures a wider range of relata, variables, influences, and events.

Finally, we focused on online causal learning here, where information flowed in rapidly and learners had no opportunity to replay and revise. However, it is possible that people are capable of reasoning more normatively in offline learning tasks when they are provided with information summarized in a timeline and can take as long as they like to consider the fit between the data and different causal hypotheses (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018). Furthermore, to the extent that summary-statistic based inference and normative inference deviate, it seems likely that people's judgments after additional thinking time could differ from their more instinctive or gut responses (Ludwin-Peery et al., 2020). Reflective thinking has been studied for decades in human reasoning and decision making (Kahneman, 2011; Sloman, 1996), while it is less studied in causal inference. The normative vs. summary-statistic contrast in this paper provides a potential paradigm for operationalizing the role of reflective thinking in causal inference.

5.8.5 Conclusions

In this paper, we showed that people can use information in continuous real time to learn about causal systems that potentially contain generative and preventative causal relationships. Their performance was influenced by multiple factors, including the nature of the causal influences (generative, non-causal, preventative), interactions with neighboring connections, base rate regularity, and intervention patterns. We laid out both a normative framework and a process-level model. Both qualitatively and quantitatively, human judgments were better captured by the process-level summary-statistic account, capturing the idea that people may infer causal structure via statistical cues such as average delays and counts that are much easier to track in real time than the exact generative model likelihoods. This work thus provides a quantitative account of how people manage to learn causal structure, in particular preventative influences, on the basis of continuous temporal dynamics. This contributes to our understanding of natural cognition and sheds light on the challenging question of how any cognitive agent can succeed in forming an internal causal model of a complex and continuous environment.

Chapter 6

Active causal structure learning in continuous time

> "The ability to know one's limitations, to recognize the bounds of one's own comprehension — this is a kind of knowing that approaches wisdom."
>
> *Leah Hager Cohen*

In Chapter 5 we have seen how people passively learn causal structures by observing predetermined temporal evidence. In this chapter, I will investigate how people *actively* learn in continuous time. That is, how they intervene in causal systems to collect evidence by themselves, and use this evidence to learn causal structures. I will focus on when and where they intervene and how this shapes their learning. Across two experiments, it is found that participants' accuracy depends on both the informativeness and evidential complexity of the data they generate. Moreover, participants' intervention choices strike a balance between maximizing expected information and minimizing inferential complexity. People time and target their interventions to create simple yet informative causal dynamics. I discuss how the continuous-time setting challenges existing computational accounts of active causal learning, and argue that metacognitive awareness of one's inferential limitations plays a critical role for successful learning in the wild.

Content from this chapter is a reprint of the material as it appears in Gong et al. (2023).

6.1 Introduction

The ability to predict, plan, and control events in the world demands a sophisticated representation of the world's causal structure. Learning such a causal model requires gathering causal

evidence through *interventions* (Pearl, 2000) — actions that manipulate the environment in ways that reveal what causes what and distinguish spurious correlations from genuine causal relationships. However, learning causal structure in general, and selecting interventions in particular, are computationally challenging problems even under idealized conditions (Bramley, Dayan, et al., 2017). In everyday life, this challenge is compounded by the need to interact with the causal environment in real time, bringing computational constraints to the fore (Griffiths et al., 2015; Simon, 1982). In this paper, we explore how people actively learn about causal structure in real time. To do this, we introduce a causal learning task in which participants interact with causal devices in real time, deciding *when and where* to intervene in order to gather information about how the device works. To motivate our novel experiments and modeling, we first summarize prior empirical work on active causal learning and point out some of its limitations. We then introduce notions of resource-rational behavior (Lieder & Griffiths, 2020; Simon, 1982) that serve as a guideline for our computational modeling framework. We then investigate human active learning about a range of acyclic and cyclic causal devices in two experiments. We analyze participants' causal judgments and intervention patterns both descriptively and through comparison with a range of models. We contrast an unbounded computational account that optimizes expected information density of its interactions with the devices with bounded models that balance information and inferential complexity. Finally, we discuss the broader implications of this perspective on accounts of human learning.

6.1.1 Prior work on active causal learning

Everyday cognition is rich with causal beliefs that explain the progression of events, shape our predictions about what is to come, and allow us to choose actions to realize our goals. For example, you might recognize a squeaking sound as caused by the opening of your garden gate, predict the doorbell will ring with your food delivery and get up to answer the door in anticipation. Many researchers have used a causal Bayesian network framework to study how people build up and represent networks of beliefs about causal mechanisms and affordances (Lagnado & Sloman, 2002; Sobel & Kushnir, 2006; Steyvers et al., 2003; Griffiths & Tenenbaum, 2009; L. E. Schulz et al., 2007; Bramley, Dayan, et al., 2017; Meder et al., 2010; Rehder, 2014; Lucas & Griffiths, 2010; Stephan, Tentori, et al., 2021; Rottman & Hastie, 2016). While the particulars of these studies are diverse, many share a core set of properties illustrated in Figure 6.1a. Participants are typically asked to distinguish between a set of candidate causal structures on the basis of evidence. Often this evidence takes the form of "snapshot" samples of discrete variables' states. Most often, the variables of interest are binary with one value construed as a variable being "present" or "active", and the other state as "absent", or "inactive".

However, when only covariation data of a set of variables is available, observational samples are insufficient to uniquely reveal structure (Pearl, 2000; Spirtes et al., 2000). For example, if a

learner observes two variables co-occurring, such that when one is active (or inactive) the other one tends to be active (or inactive) too, they cannot tell if one is causing the other or if they share an unobserved common cause. One solution is to intervene (Pearl, 2000) — manipulating one or more variables in the system by fixing them to particular values and observing how this affects the rest of the system. A number of experiments have allowed participants to perform such interventions in order to support their learning (Bramley, Dayan, et al., 2017; Bramley et al., 2015; Steyvers et al., 2003; Lagnado & Sloman, 2002; Coenen et al., 2015).

Studies have shown that well chosen interventions can speed up learning, allowing learners to target their uncertainty and quickly narrow in on the true model. However, poorly chosen interventions can be worse than random actions or passive observations (Settles, 2009). In the covariation-data setting, adults and children have been found to be able to select informative interventions and learn successfully from them about probabilistic systems involving a handful of variables (Bramley et al., 2015; Coenen et al., 2015; Steyvers et al., 2003; McCormack et al., 2016; Meng et al., 2018). At a normative level, informative interventions are those whose consequences are expected to strongly distinguish among the potential hypotheses, maximally decreasing global uncertainty in expectation (Tong & Koller, 2001), or maximizing the chances of inferring the true causal structure that gave rise to the data (Nelson, 2005). A number of experiments have demonstrated broad alignment with these norms in both adults and children, but also departures from the normative predictions which suggest that process-level considerations are important for fully characterizing people's inferences (Bramley, Dayan, et al., 2017; Bramley et al., 2015). For example, people often chose an intervention that is expected to confirm or refute a currently-favored hypothesis rather than one that provides more information about the full hypothesis space (Coenen et al., 2015; Meng et al., 2018; Steyvers et al., 2003; Klayman & Ha, 1989). People sometimes also rely on generic strategies such as systematically fixing the values of some variables while varying others in order to isolate one potential relationship at a time (Bramley et al., 2015; McCormack et al., 2016; L. E. Schulz et al., 2007). One manifestation of this is the so-called *control of variables* strategy in which a set of candidate causal variables are fixed and one variable is changed in each experiment (Zimmerman, 2007; Kuhn & Brannock, 1977; Chen & Klahr, 1999). Following such a strategy has been emphasized in developmental psychology as a marker of mature scientific experimentation, but this strategy turns out to be suboptimal in certain environments (Coenen, Ruggeri, et al., 2019; Bramley et al., 2022). Finally, adults choose interventions adaptively, taking into account environmental factors, such as time pressure, as well as whether a strategy was informative in the past (Coenen et al., 2015).

6.1.2 What prior work has neglected

Previous work on causal learning has largely focused on situations that mimic idealized laboratory conditions. In these studies, participants perform interventions in a discrete trial-by-trial manner,

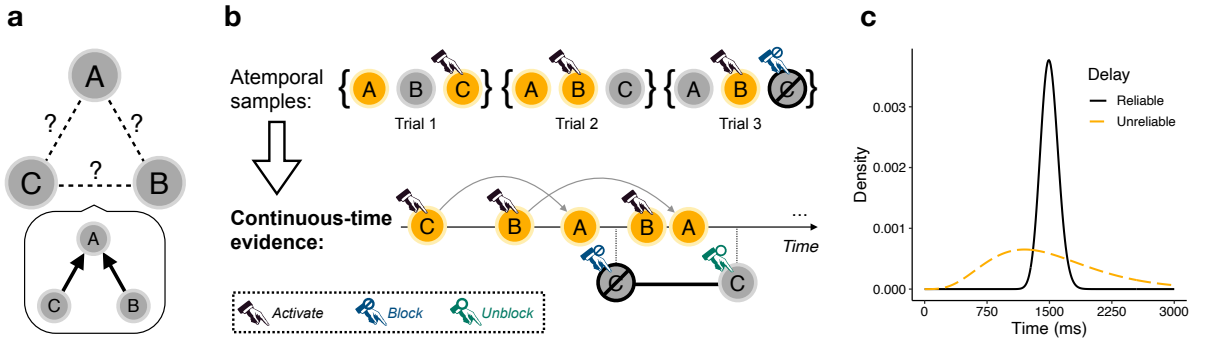


Figure 6.1: Illustration of causal systems and sample types. a) Three components with causal connections unknown to the learner. b) Atemporal samples under three trials, and its possible corresponding continuous-time samples. Yellow indicates a component activated. In the continuous-time setting, interventions activate components in real time and effects may occur intermingled on the timeline. Arrows indicate the underlying generative process unbeknownst to the learner. c) Gamma density distributions under reliable vs. unreliable causal delays in the current experiments. Both probability distributions have a mean of 1.5 s with different standard deviations (0.1 s for reliable and 0.7 s for unreliable delays).

and the values of all variables are revealed all at once. In this way, participants are invited to generate and reason from a series of independent observations. Figure 6.1b illustrates a example of this *atemporal* evidence, generated from interventions and subsequent observations of the variables in a stochastic system (see Coenen et al., 2015; Bramley et al., 2015, for example). Information arrives in three independent trials in the form of variable states (yellow = present or active; gray = absent or inactive) conditional on interventions (i.e. variables fixed on or blocked off by the learner).

At a computational level, the problem is one of identifying the true generative causal Bayesian network — the parameterized graph that captures the patterns of covariation between the variables under both observations and any hypothetical intervention (Pearl, 2000). For example, in Trial 1 in Figure 6.1b we see that, conditional on an intervention that activates C, A activated and B did not. This can be written as $\{A = 1, B = 0 | \text{ Do}[C = 1]\}$, where 1 indicates a variable was active, 0 indicates it was inactive and Do[.] indicates a variable was fixed through intervention and thereby disconnected from its normal causes on this trial. Interventions can target multiple variables. For example, in Trial 3, both B and C are manipulated as B is activated and C is blocked (fixed to be inactivate). In this kind of task, ideal inference and intervention selection are well-understood computationally, facilitating comparison between behavior and rational norms (e.g. Rottman & Hastie, 2014). However, this task setup differs in several respects from the causal learning and reasoning problems people face in daily life when they (1) take into account temporal information, (2) deal with evidence that is interdependent, and (3) encounter causal learning problems when the underlying causal mechanism may be cyclic.

Time

Most previous studies removed temporal information, including the order of events and the delays between them. For example, Coenen et al. (2015) described a cover story of computer-chip systems where the causal relationships are the passage of electrical current from the energy source to the components, occurring too fast to distinguish order of activation. Other studies only allowed participants to view the final outcome (Bramley et al., 2015; Rottman & Keil, 2012). In contrast, many everyday causal relationships take time to propagate, meaning that the temporal order and delay between events is relevant for inferring causal relationships. The notion that causes must precede their effects is foundational to the concept of causation (Burns & McCormack, 2009; Lagnado et al., 2007; White, 2006). Indeed, people have been shown to rely on temporal order to guide causal inference even when it conflicts with covariation information (Lagnado & Sloman, 2006), and to assign low probabilities to mechanistic explanations for event sequences that would require an effect to have occurred at the same time as its cause (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018).

People not only have expectations about the order of events but also about the delays between them, giving higher causal strength ratings when delays between a putative cause and effect are short and reliable (Greville & Buehner, 2010) as well as when they conform to prior or mechanistic expectations (Hagmayer & Waldmann, 2002; Buehner & McGregor, 2006; Buehner & May, 2004). For example, Buehner & McGregor (2006) found that participants gave higher causal judgments about the insertion of a ball turning on a light on a physical apparatus when the light came on after a few seconds rather than instantly, if they were aware it took time for the ball to roll through the apparatus and reach the light switch. A separate line of work has studied inference and representation of continuous variables in continuous time (Davis et al., 2020; Soo & Rottman, 2018). However, temporal information is yet to be examined in the context of active causal learning.

Interdependence

Under the laboratory conditions created in prior experiments, evidence is taken to come from multiple "independent, identically distributed" (i.i.d.) observations or interventions. For example, trials may pertain to different test subjects drawn from the same population (e.g. pairs of patients and treatments; Buehner et al., 2003), or might involve repeated interactions with the same causal mechanism, but collected via a protocol that ensures variables "reset" from one trial to the next (e.g. the "blicket detector"; Gopnik et al., 2001; Lucas et al., 2014). However, it is rare for everyday experience to exhibit these properties. In life, there is no magic reset button. It is hard to be sure whether and when a causal system has been reset without understanding its underlying mechanism (defeating the goal of the exercise).

To illustrate this, imagine wondering why your puppy is unusually excited one evening. You consider two candidate hypotheses: Perhaps his elevated mood is due to a new variety of dog food you fed him at 5pm, or perhaps it is because of a new floral scent on the road where you walked him at 6pm. The puppy might still be happy about his dinner even after having smelt the flowers. A poor approach to resolving the question would be to always feed him beside the flower bed. It would be better to vary the relative time of walking and feeding him while keeping a close eye on the time intervals implied by different causal explanations.

This example illustrates that active learning in everyday life is better understood as a rolling sequence of interventions, with cause and effect events unfolding on a single continuous timeline. In fact, Rottman & Keil (2012) found that when presented with a sequence of experimental results, even paired with a cover story that implied these experiments were independent, many participants judged causal relationships by how values changed relative to their state on the preceding observation, rather than treating the samples as independent (see also Derringer & Rottman, 2018). This suggests that when evidence arrives over time, people strongly assume temporal dependence. Thus, it seems that temporal dependence not only reflects genuine causal phenomena but that it may also better match laypeople's intuitive causal theories than time-agnostic Bayesian networks do.

Causal cycles

Causal learning studies have largely focused on acyclic causal systems where causal influences flow only in one direction, never revisiting the same components. This is partly due to the conceptual and mathematical convenience afforded by the formalism of acyclic causal Bayesian networks (see Rottman & Hastie, 2014, for a review). The continuous-time setting enables us to investigate cyclic causal relationships. A causal mechanism is cyclic if it has at least one component whose descendants include itself (Pearl, 2000). This means that the components that form part of the cycle, or outputs from it, may occur in repeated alternating fashion (e.g. a bidirectional connection $A \leftrightarrow B$ could generate a sequence of events $A, B, A, B, A, ...$). Many causal processes in the natural world are cyclic (Malthus, 1872), and people frequently report causal beliefs that include cyclic relationships when allowed to do so in experiments (Kim & Ahn, 2002; Nikolic & Lagnado, 2015; Sloman et al., 1998; Rehder, 2017), making this an important aspect of causal cognition to study.

6.1.3 The current paradigm

The learning problem Departing from the atemporal setting, we focus on what people can learn from interventions and observations of *events* within a single continuous timeline. We study a setting in which effect events follow their causes with some stochastic but predictable delay.

This causally-connected-point-event setting has been used in a number of recent studies of temporal causal reasoning. It rests on a firm mathematical foundation that supports normative inferences from temporal information to causal structure. Griffiths & Tenenbaum (2009); Greville & Buehner (2010) first demonstrated that people can infer how pairs of variables affect one another from observing sequences of point events. Pacer & Griffiths (2012, 2015) developed a model that infers causal relationships based on the occurrence of putative cause events that influence the *rate* at which the relevant effect events occur over time. Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018) built models that combined hard order constraints with soft delay expectations to best capture structure judgments: Even when order information was fixed, participants were still sensitive to the variation in inter-activation delays between events and used it to distinguish between certain causal structures.

More recently, research has focused on so-called "actual causation" (Halpern, 2016): The question of which out of multiple candidate events actually brought about the outcome (Stephan et al., 2020; Gerstenberg et al., 2021). Using key ideas from the "actual causation" literature, recent work has looked at how causal structures can be identified from temporal information by considering the different possible causal pathways that could have produced the observed events conditional on different underlying causal mechanisms (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Gong & Bramley, 2020; Valentin et al., 2020). We follow this approach, using Gamma distributions to model the distribution of causal delays exhibited by a particular causal component of a device across instances (see Figure 6.1b). Gamma distributions define a probability density over $(0, +\infty)$ via a shape parameter α, and rate parameter β allowing for a variety of causal delay distributions with differing means and more or less variability (see Figure 6.1c).

While temporal information was key to how the evidence was presented in the studies above, the data was not fully continuous in the sense in participants experiences were still broken into separate independent episodes. For example, in order to set things in motion in Bramley, Gerstenberg, Mayrhofer, & Lagnado (2018), each clip began with the system at rest perturbed by an exogenously caused root-component activation, with effects following from there. Since components could only activate once in these tasks, the system would quickly reach a steady state. This still departs from a fully continuous-time setting in which interventions and effects are intermingled and components may exhibit multiple activations within the same episode. A fully continuous setting makes it more difficult to figure out what caused what because any given event might be attributed to an earlier-occurring event and might have its own effects that are still to occur.

Activating and blocking interventions in time In our experiments we will allow learners to intervene on the causal system in two ways:

1. By *activating* components, thus potentially setting in motion a new sequence of events.

2. By *blocking* components, thus preventing that component both from being activated and from activating any other components until it is unblocked again.

Activating and blocking are superficially analogous to fixing variables to be on Do$[X = 1]$ or off Do$[X = 0]$ in the atemporal setting. However, they also differ in important ways. In the continuous-time setting, activating does not disconnect a component from its normal causes. The intervened-on component can be activated again an arbitrary number of times during the same episode either by the intervener or when caused by other variables in the system. For instance, if the activated component is part of a cycle, we would expect it to be re-activated repeatedly following its initial activation until one of the causal connections fails. Thus, activation is better thought of as a shock to the system than as a form of graph surgery.

On the other hand, *blocking* actions do exemplify the "graph surgery" property in the sense of Pearl (2000). They disconnect the blocked component from its normal causes until it is unblocked again (Figure 6.1b). In the atemporal setting, blocking is essential for discriminating between certain structures (Bramley et al., 2015; McCormack et al., 2016; L. E. Schulz et al., 2007). For example, turning on a single component (i.e. Do$[A = 1]$, Do$[B = 1]$ or Do$[C = 1]$) generates a similar pattern of dependence under a $A \rightarrow B \rightarrow C$ chain and a $C \leftarrow A \rightarrow B \rightarrow C$ fully connected structure — activating A affects B and C, activating B affects C but not A, activating C neither affects A nor B. This makes it difficult and inefficient to distinguish these structures based on activating interventions alone and impossible in deterministic settings. To identify whether there is direct link between A and C, one must turn on A while simultaneously blocking or disabling B (i.e. $Do[A = 1, B = 0]$).

The current continuous-time setting endows blocking with different implications. Since causes generate effects individually, blocking is not strictly required to distinguish direct and indirect paths. Going back to the chain vs. fully connected example above, a fully connected system would normally produce two staggered activations of C following an intervention on A while the chain would produce only one, making them distinguishable in principle. Nevertheless, blocking may be useful for reducing computational complexity and ambiguity of parsing the consequent event sequences. The learner can use blocking to reduce the event numbers, or remove a component from consideration, while still making remaining events informative (see the section below for two examples).

6.1.4 Cognitive resource limitations

In the atemporal setting, causal reasoning is a little like crafting an essay: Evidence can be collected, organized and put together carefully with room for reorganizing and backtracking in searching for an effective structure. However, real-time learning more closely resembles the problem of writing under exam conditions: One must react immediately to the prompts, bringing ones inferential tools to bear quickly and efficiently without the luxury of time to backtrack.

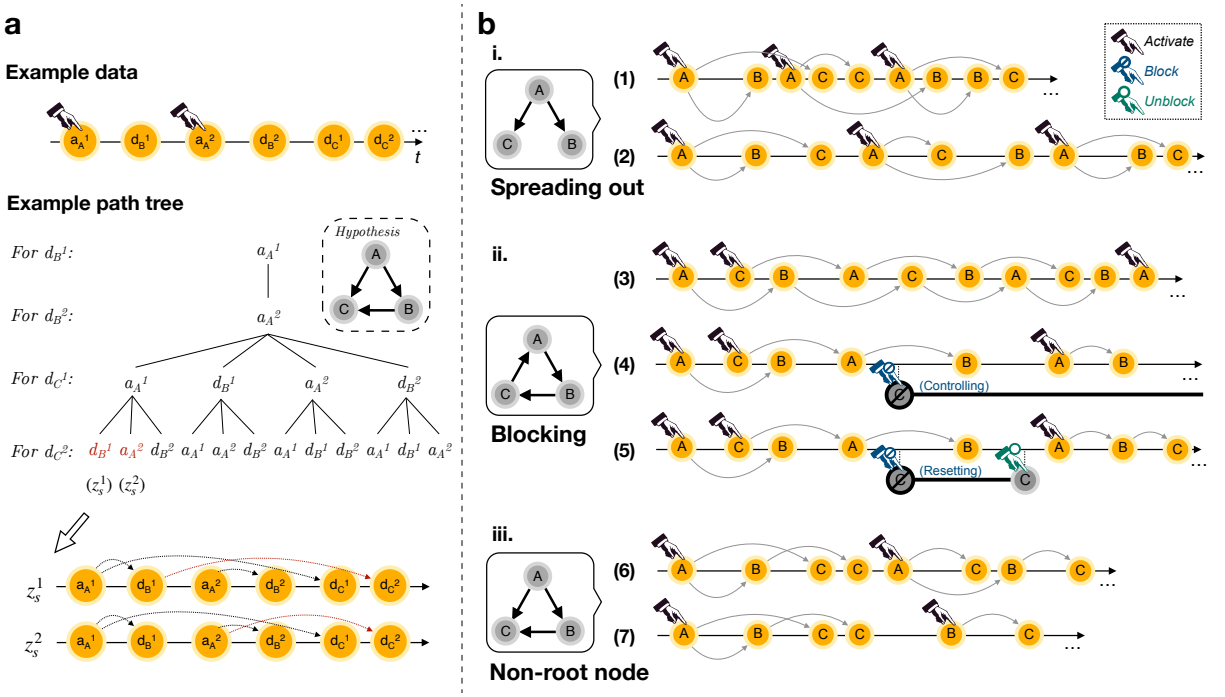


Figure 6.2: Sketch of ideal observer inference algorithm and approaches to minimizing complexity. a) Ideal Bayesian inference considers each possible structure hypothesis s and every possible causal path z_s that could describe how that structure produced the observations. The number of possible paths grows rapidly in the number of "nearby" events as illustrated with an example recursion tree showing all twelve paths connecting events $a_A^1 ... d_C^2$ conditional on the structure $C \leftarrow A \rightarrow B \rightarrow C$ ($a^{(\text{index})}_{\text{component}}$: activating interventions; $d^{(\text{index})}_{\text{component}}$: effect events, see Appendix B.1 for a full description of notation). Two paths z_s^1 and z_s^2 were further displayed in a timeline format with arrows showing the hypothesized generative process and red arrows in particular highlighting the different delay implications. b) Three examples of interventional strategies that help reduce the inferential cost of processing generated evidence. Sequences (2), (4), (5), (7) are less complex to process than Evidence (1), (3), (6).

We lay out the how an ideal Bayesian observer learns the causal structure with temporal evidence in Appendix B.1. This shows that the amount of computation needed to process the evidence compounds rapidly as more events occur. An ideal learner needs to consider all the plausible pathways through which a particular causal structure might have produced an observed pattern of events (Halpern, 2016). For example, consider intervening once each on A and B and then observing two subsequent activations of C. To calculate the overall likelihood that a "collider" structure $A \rightarrow C \leftarrow B$ could have produced this pattern, we would need to take into account two possibilities (1) that A produced the first activation of C and B the second one, or (2) that B caused the first one, and A the second one. However, there will generally be far more than two such possibilities. Figure 6.2a shows a more complex example in which twelve possible causal paths could link six events for a single causal structure. For a handful more events the number of paths can easily grow into the millions.

Challenging such a naive idealized account of causal inference is the basic fact that human minds are bounded in their capacity to compute and store information. Human reasoning and decision-making necessarily deviates from such intractable, computational-level ideals (Anderson, 1990; Simon, 1982). Given the computational complexity and conceptual centrality of structure learning in cognition, we expect computational costs to play a large role when structure inference must take place in real time (Christiansen & Chater, 2016). Several process-level proposals have been explored in the literature as candidates for how people approximate normative structure inference. People may only consider a few sampled hypotheses (Bonawitz et al., 2014), incrementally adjust a focal hypothesis to accommodate new evidence (Bramley, Dayan, et al., 2017; Davis et al., 2020; Fernbach & Sloman, 2009), rely on recent evidence (Bramley et al., 2015; Bramley, Dayan, et al., 2017), rely on summary statistics (Gong & Bramley, 2020, 2023a; Ullman et al., 2018), or on simple heuristics such as equating temporal order with causal order (Burns & McCormack, 2009; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Bramley, Mayrhofer, et al., 2017).

While we expect some combination of the above ideas to be in play in how participants solve our task, we here explore a complementary facet of causal learning, the active gathering of evidence through interventions to support the inference process. We ask whether people time and target their interventions so as to manage the inferential complexity of parsing resultant evidence, while still producing informative evidence overall. Figure 6.2b illustrates this idea by displaying three potential intervention strategies for managing evidential complexity. In the first example, the ground truth is $B \leftarrow A \rightarrow C$. An unbounded ideal learner learns about as much from Evidence 1 as from Evidence 2. However, if we assume that the ability to process evidence is a function of its complexity and that this is related to the density of the events being reasoned about, then it is clear that Evidence 2 is the more useful for a bounded learner. Here, the events are better separated, and so there is much less ambiguity about the plausible causes of each token event, therefore less need to engage in costly averaging over many potential causal pathways under each structure hypothesis.

Bounded learners may also choose to *block* components of a system to make the event stream manageable for reasoning. As shown in Evidence 3 of Figure 6.2b, having performed two activating interventions in a cyclic system, a learner may experience a confusing pattern of parallel excitation. Evidence 4 shows how this can be avoided using a "controlled" testing strategy that blocks a component before activating another. This approach allows a learner to isolate a subsystem of a larger system. This controlled test means fewer interpretations of the evidence need be considered. For instance, in Evidence 3 remaining events are straightforwardly indicative of the substructure linking the unblocked components A and B. Well-timed blocks might also be used to impose pseudo-independence and trial-like structure within a continuous interaction. For example, as shown in Evidence 5, one might block, wait, then unblock components to "reset" a system, preventing any ongoing activity from complicating the inference process going forward.

Finally, although activating a suspected root component (here A) tends to produce more evidence about a causal structure than activating its suspected tail nodes (Evidence 6), bounded learners might sometimes avoid root components. For instance, if primarily interested in understanding a presumed downstream subpart of a complicated causal mechanism, one might intervene locally to avoid extraneous events and activity (Evidence 7). Note that those considerations are not independent. Spreading out interventions in time, for example, requires the learner to wait for the system to calm down. The same could be accomplished by blocking-and-unblocking a component to reset the system.

Finding ways to balance informativeness and complexity in generating evidence is conceptually related to the notion of bounded rationality (Simon, 1982; Anderson, 1990). The basic idea is that human minds have evolved or discovered solutions that trade off efficiently between the costs of computation and its rewards in greater accuracy or performance. In particular, one can incorporate computational costs into a solution space formally with a *resource rationality analysis* (see Lieder & Griffiths, 2020; Griffiths et al., 2015; Shenhav et al., 2017, for review). This has suggested that a number of decision making phenomena classically seen as irrational — including as anchoring and probability matching — may instead represent efficient solutions to a computation–value trade-off under some sensible approximation scheme (Callaway et al., 2022; Lieder, Griffiths, Huys, & Goodman, 2018; Hawkins et al., 2021; Dasgupta et al., 2017; Lai & Gershman, 2021). We similarly use a resource-rationality framework to analyze adults intervention choices and judgments in our tasks. This involves first considering the impact of both information and complexity on inferential success, and second, modeling intervention selection as driven by a goal of maximizing the expected informativeness of the evidence while minimizing the expected inferential cost of processing the evidence.

6.1.5 Overview of experiments

We conducted two experiments to test how people actively learn causal structures in a continuous-time setting. In both experiments, we manipulated the reliability of the cause-effect delays and included a range of acyclic and cyclic causal structures. In Experiment 1, we only allowed participants to activate components while in Experiment 2, we also allowed them to block components.

In line with our normative account of causal inference in this setting, we hypothesized that performance would be lower in the irregular delay condition given that evidence about what caused what is more ambiguous. In line with our bounded inference account, we also hypothesized that performance would be worse in cyclic systems given the likely increase in event density, interdependence and concomitant complexity. However, we further expected accuracy to depend on the quality and reactivity of participants' intervention choices. Thus, we also examine whether and how participants' intervention selection differs across devices and delay conditions, asking to what extent intervention choice is reactive to the behavior of the device being explored, and whether

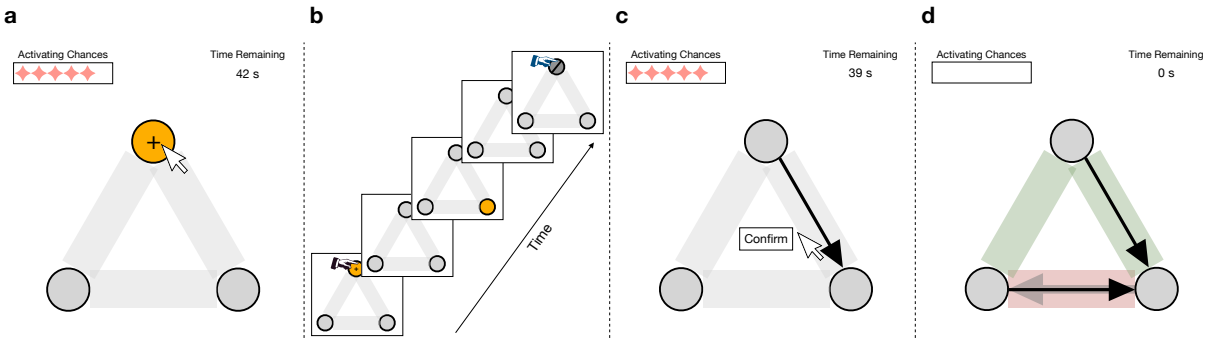


Figure 6.3: Experimental procedure. a) Experimental interface. Up to 6 interventions could be performed by clicking on the components during the 45 second trial. b) Example timeline. Interventions lead to subsequent activations determined by the direction and delay of the causal connections in the true model. c) Judgments. Participants can indicate their beliefs about the structure during the trials by clicking on the edges. Participants in Experiment 1 needed to click the confirm button to lock their answers. d) Feedback. At the end of each trial feedback was provided (green = correct; red = incorrect; wide gray arrows in the background indicate ground truth).

this reactivity reflects rational anticipation and active management of expected information gain and evidential complexity.

6.2 Experiment 1: Causal structure induction in continuous time

6.2.1 Methods

Participants

Seventy-four participants (40 female, 34 male, aged 30 ± 11) were recruited from Prolific Academic and were randomly assigned to either the reliable-delay ($N = 36$) or unreliable-delay ($N = 38$) condition. Participants received a basic payment of £1 and a bonus depending on performance (see *Incentives* section). Nine additional participants were tested but removed from the analysis because they left the default "unconnected" connection judgment for all causal component pairs for all trials ($N = 6$) or had at least one trial in which they performed no interventions at all ($N = 3$). The sample size was chosen to be in line with related work on causal learning (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Coenen, Ruggeri, et al., 2019).[1]

[1]The experimental procedure, data, and analysis code are available at: `https://github.com/tianweigong/time_and_intervention`.

Design

Participants were randomly assigned to one of two delay conditions: (1) reliable delays ($M \pm SD = 1.5 \pm 0.1$ s, i.e. $\alpha = 200, \beta = \alpha/1500$ in the Gamma distribution) or (2) unreliable delays (1.5 ± 0.7 s, i.e. $\alpha = 5, \beta = \alpha/1500$, see Figure 6.1c).

Participants were asked to investigate abstract causal "devices" connected by hidden causal links (Figure 6.3a). The causal links produce point events in the form of activations of the device's components over time. For causally related components, an activated component will probabilistically activate each of its effect components once after some delay. All causal connections worked 90% of the time and no events occurred without being caused by an intervention or other event (i.e. none of the components activated spontaneously). Participants were informed and tested on this in the instructions.

Each participant learned about 12 test devices with either 3 or 4 components, including 6 acyclic structures and 6 cyclic structures (Figure 6.4). The acyclic structures were chosen to exemplify a variety of causal relationships including common effects, i.e. "colliders" (`Acyclic1` and `Acyclic4`), chains (`Acyclic2` and `Acyclic5`), "forks" (`Acyclic3` and `Acyclic6`). The cyclic devices were chosen so as to approximately match the number of edges in the acyclic systems while investigating a variety of arrangements. These included full loops (`Cyclic1` and `Cyclic4`) and short loops with incoming connections (`Cyclic2` and `Cyclic5`) and outgoing connections (`Cyclic3` and `Cyclic6`).

Interface Figure 6.3 shows the task interface. For each device, participants saw the 3- or 4-components visualized as gray circles evenly spaced on a white background. Participants had 45 seconds to learn about how the components were connected. During this time, they could intervene and activate components by left-clicking on them up to 6 times. Intervened-on components were marked by a "+" symbol (Figure 6.3a). All activated components turned yellow for 200 ms and then returned to gray (Figure 6.3b). At the beginning, all components were inactive, and no connecting links were marked between them.

Participants were able to indicate their current belief about the causal structure as often as they liked during each learning problem. To do so, participants clicked on the gray area between components to toggle between a causal connection in either direction, both directions, or no connection. Each click cycled through the options ($A \rightarrow B$, $B \leftarrow A$, $A \leftrightarrow B$, no relationship) in a random order varied between participants. Participants confirmed their choices by clicking a confirm button that appeared in the middle (Figure 6.3c). Links did not disappear after being confirmed, so participants were still able to update earlier judgments. What participants had marked at the end of 45s was automatically registered as the final judgment for that trial. At the end of the trial, participants received feedback about which connections they had marked correctly or incorrectly (Figure 6.3d). Since any pair of components might be unconnected, have

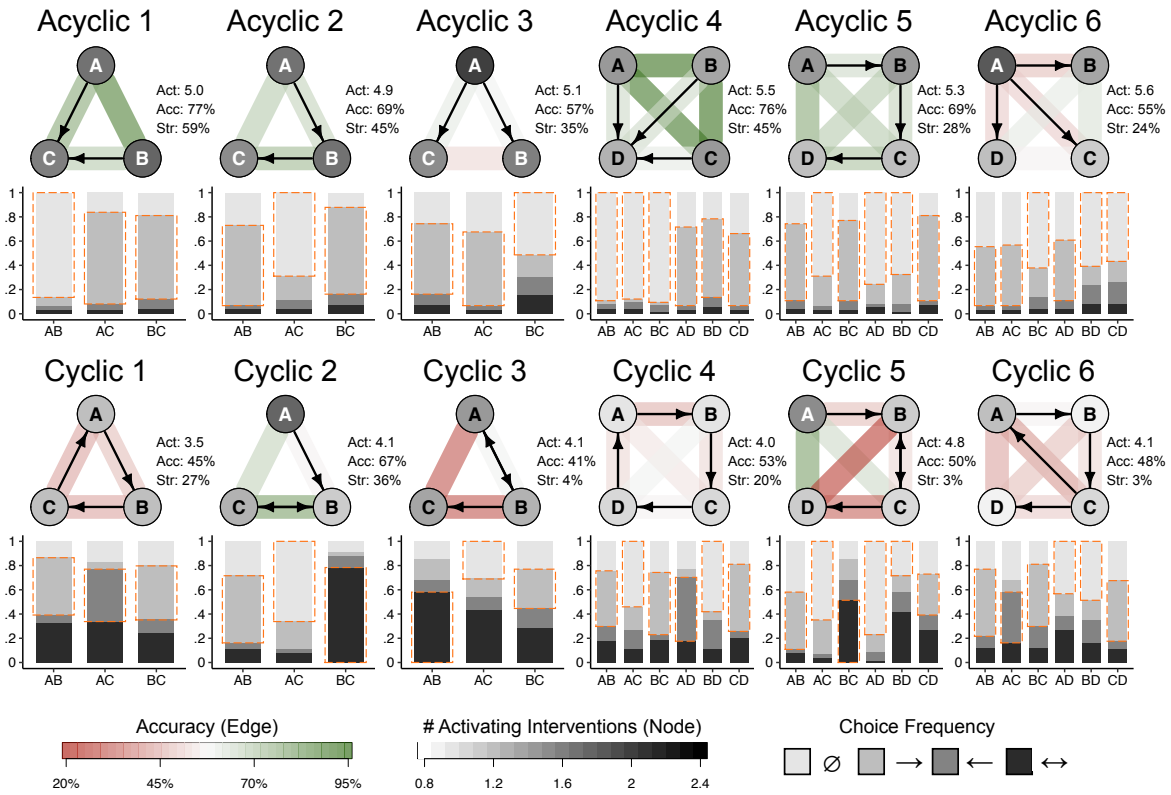


Figure 6.4: Causal link identification and activating intervention choices in Experiment 1. Color edge shading indicates accuracy. Node shading indicates activating intervention choice prevalence by component. Bar plots show the proportion of different choices on each link (e.g. for AB, "$\varnothing$" means "no connection between A and B"; "$\rightarrow$" means "$A \rightarrow B$"; "$\leftarrow$" means "$A \leftarrow B$"; "$\leftrightarrow$" means "$A \leftrightarrow B$") with orange used to highlight the ground truth. Note: Act = average number of activating interventions performed; Acc = mean accuracy; Str = proportion of participants who detected the whole structure correctly.

a directed ($A \rightarrow B$ or $B \rightarrow A$), or bidirectional ($A \leftrightarrow B$) causal connection, the response space includes 64 possible structures for 3-variable devices and 4096 possible structures for 4-variable devices of which exactly one truly reflects the hidden causal structure.

Incentives We incentivized participants to mark the correct causal links as early as possible within the trial by rewarding them based on their accuracy at a random time point during each problem. This bonus scheme means it is in participants' interest to register their best guess accurately and early, and to update it whenever their conclusions change during a learning episode (cf. Bramley, Dayan, et al., 2017). The scheme also shapes the nature of an ideal intervention strategy, meaning one should balance the benefits for intervention selection of waiting until one knows more, against the opportunity cost of waiting too long and missing out on what could have been learned from an earlier intervention. In order to perform well in our task, learners need not only consider *how* and *where* to intervene next, but also *when* to do so.

Table 6.1: Accuracy separated by conditions.

	Delay Reliability		Structure Cyclicity			Structure Nodes	
	Reliable	Unreliable	Unlinked	Acyclic	Cyclic	3-node	4-node
Experiment 1	62%±33%	55%±32%	–	67%±33%	51%±31%	59%±33%	59%±30%
Experiment 2	61%±35%	61%±34%	77%±39%	65%±33%	52%±32%	60%±37%	61%±31%

Procedure

In the main task, each participant faced 12 test devices in random order with randomly positioned and unlabeled components. Prior to the inference task, participants completed instructions, a practice trial and comprehension checks. Participants were told that they would be investigating the causal structure of a number of abstract "devices". In the instructions, participants were trained on the true cause–effect delays in their condition and shown a video example of a device with its causal links revealed. They were then trained on how to provide structure judgments. Participants learned that they would receive a £0.03 bonus for each connection correctly marked at a randomly chosen and unmarked point during each trial (for a theoretical maximum total bonus of £1.62). This was emphasized in the instructions to encourage participants to mark connections as quickly as possible. Participants had to correctly answer 5 comprehension check questions before proceeding to the main task. Finally, participants completed a practice trial on a device with a collider structure (`Acyclic1` in Figure 6.4).

6.2.2 Results

We first report participants' judgment accuracy (i.e. what proportion of connections participants correctly identify at the end of each trial) by delay condition (reliable vs. unreliable), device type (acyclic vs. cyclic), and number of components (3 vs. 4). We then discuss characteristic error patterns under specific causal structures. Our accuracy analyses use linear mixed-effect models (LMMs) including random slopes and intercepts for subject ID and structure type (Brauer & Curtin, 2018). For all LMMs we report standard coefficient estimates βs (that show how many units of standard deviations the outcome variable changes when the independent variable changes from one condition to the other), t values, significance, and 95% confidence intervals (CI). We then compare participants' trial-by-trial accuracy against the predictions of a normative inference model. We explore whether deviations from normative responding are related to the density of events in that trial as a basic index of complexity. We will then focus on participants' intervention choices and explore whether interventions are driven by a trade-off between the expected evidence strength and complexity.

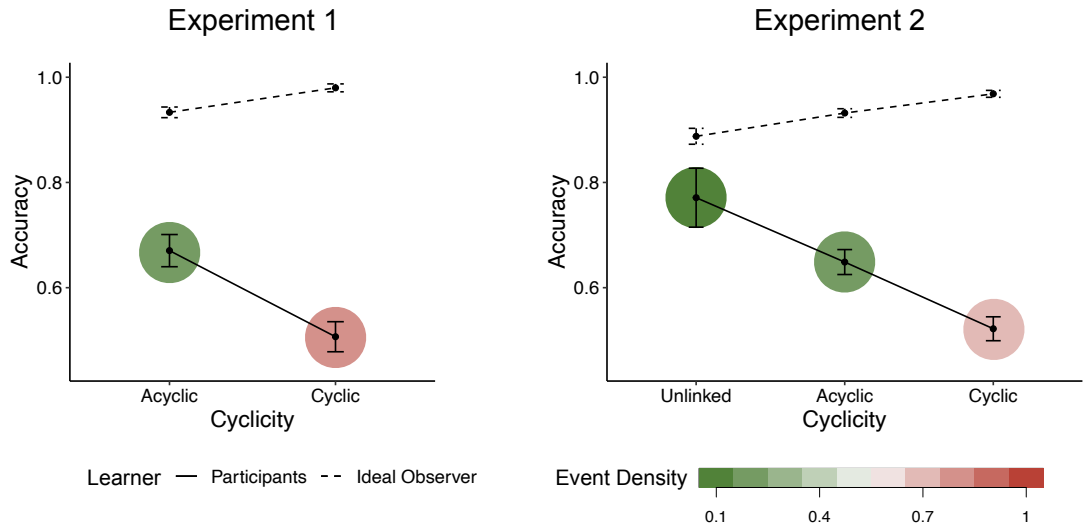


Figure 6.5: Participants vs. the ideal observer's accuracy and event density upon human generated evidence. Error bars indicate 95% confidence intervals.

Accuracy Participants confirmed their causal judgments 2.45 ± 1.31 times per trial. Final judgments — i.e. what participants had marked at the end of the trial — identified the majority of the causal connections correctly ($62\% \pm 34\%$) but with marked variation across and within devices. Participants' final judgments generally improved on the accuracy of their initial judgments — i.e. what participants had marked as their first answers — in the 79% of trials in which participants made more than one judgment (initial accuracy: $58\% \pm 30\%$, $\beta = 0.11$, $t = 2.36$, $p = .024$, $CI = [0.02, 0.20]$). In the following, we focus on participants' final structure judgments.

Participants performed significantly above chance in both delay groups (chance: 25%, reliable: $t(35) = 10.83$, $p < .001$, Cohen's $d = 1.80$; unreliable: $t(37) = 11.44$, $p < .001$, Cohen's $d = 1.86$) and were above chance for all 12 structures taken individually in both the reliable ($ts(35) > 4.15$, $ps < .001$) and the unreliable condition ($ts(37) > 3.49$, $ps < .01$) with the exception of `Cyclic1` (unreliable: $t(37) = 1.94$, $p = .06$, Figure 6.4). Table 6.1 shows accuracy separated by condition. There was a main effect of structure cyclicity ($\beta = 0.50$, $t = 2.69$, $p = .026$, $CI = [0.14, 0.86]$) such that the accuracy was higher for acyclic than cyclic structures. There was no main effect of delay reliability ($\beta = 0.21$, $t = 1.47$, $p = .15$, $CI = [-0.07, 0.49]$), or of number of components ($\beta = 0.02$, $t = 0.09$, $p = .93$, $CI = [-0.34, 0.37]$). Nor were there any two- or three-way interactions ($ps > .10$).

Error patterns Participants were best at inferring the structure of colliders (`Acyclic1` and `Acyclic4`, see Figure 6.4). These structures were naturally simple in their evidence since no intervention would cause more than one effect. For the three-component chain (`Acyclic2`), 15% of participants added an erroneous additional direct connection $A \rightarrow C$. Similarly, for the four

component chain (`Acyclic5`), participants frequently also added one or more "short cut" links from $A \rightarrow C$ (12%), $A \rightarrow D$ (8%) or $B \rightarrow D$ (13%) in addition to the true connections. These errors cohere with previous findings suggesting that people rely on local computations when inferring causal structure, resulting in the addition of extraneous connections in chain structures (Davis et al., 2020; Fernbach & Sloman, 2009). No one mistook the chain `Acyclic2` for a fork with the same root component (i.e. `Acyclic3`), however 18% mistook the fork `Acyclic3` for a chain with the same root component ($A \rightarrow B \rightarrow C$ or $A \rightarrow C \rightarrow B$). This lines up with the idea that people tend to fall-back on temporal order as a cue to causal order (McCormack et al., 2016; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018), tending to link the effect components of a fork in whatever order they happened to activate. These error patterns did not differ significantly between reliable vs. reliable groups (χ^2 tests, $ps > .10$).

For the cyclic structures, participants' judgments varied considerably so we focus on the individual-connection-level errors as shown in Figure 6.4. In the full loops (`Cyclic1` and `Cyclic4`), participants frequently judged directed or disconnected links as bidirectionally connected (Figure 6.4, black bars). This was more prevalent in the unreliable group than in the reliable group (`Cyclic1`: 37% vs. 23%, $\chi^2(1) = 4.31$, $p = .04$; `Cyclic4`: 22% vs. 10%, $\chi^2(1) = 11.41$, $p < .001$), suggesting that reliable delays make it easier to detect full loop structures. This makes sense since regular delays produce much more sequentially reliable and predictable patterns of reactivation. Participants had relatively little trouble at identifying loops with incoming connections (`Cyclic2`). However, performance was very poor for structures comprised of feedback loops with outgoing connections (`Cyclic3`, `Cyclic5`, and `Cyclic6`). The outgoing component (C in `Cyclic3`, D in `Cyclic5` and `Cyclic6`) was frequently taken to be a constituent of the feedback loop, often being assigned a bidirectional connection with one of the loop constituents. This is reasonable since, for these structures, recurrent and close-in-time events occurred not only at the components forming the loop themselves but also for the output components, making it difficult to tell which components were involved in actively sustaining the looping pattern of activations. Participants often connected an output component to the loop element that typically activated in close temporal proximity. For example, many participants marked a $C \leftrightarrow A$ connection in `Cyclic3`, $D \leftrightarrow B$ in `Cyclic5`, and $D \leftrightarrow A$ in `Cyclic6`, in spite of the fact that this temporal proximity is really due to them sharing a common cause.

Participants were normally correct about whether the structure was cyclic or acyclic. Participants' structure judgments belonged to the correct class 82%±38% of time for the acyclic class and 77%±42% of time for the cyclic class. There was no difference in the frequency of mistaking cyclic for acyclic vs. acyclic for cyclic ($t(73) = 1.15$, $p = 0.25$), and the rate did not differ between reliable and unreliable delay conditions ($ps > .10$).

Informativeness and event density We calculated the accuracy of an ideal observer (IO) based on the 45-seconds of evidence generated by each participant on each trial. This acts as

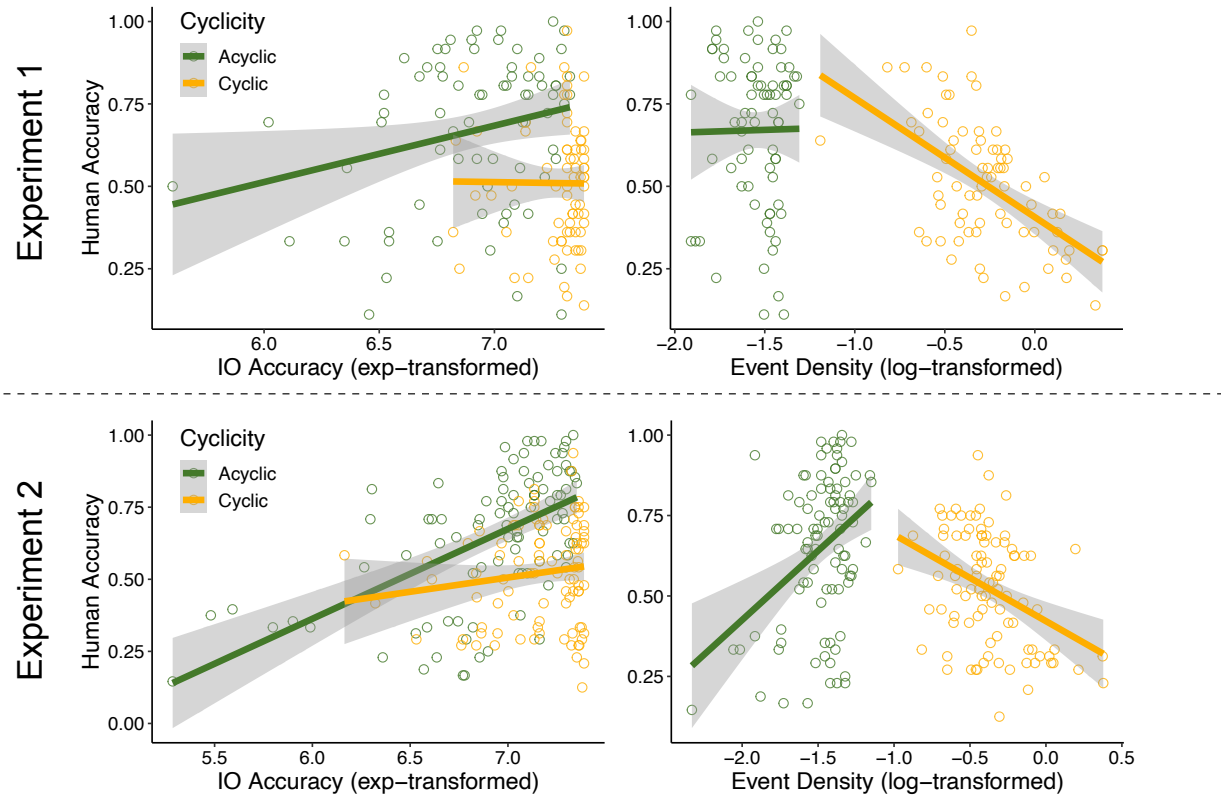


Figure 6.6: Scatterplots of evidence informativeness (IO accuracy) and event density and participants' accuracy. Each data point shows an individual's average performance for acyclic or cyclic causal structures. One data point (4.96, 0.28) under the cyclic condition was removed from the upper left panel for visualization.

a measure of how *informative* the evidence generated by the participants was (cf. Bramley et al., 2015). The IO was more accurate in the reliable (97%±7%) than the unreliable (94%±12%) condition ($\beta = 0.23$, $t = 2.61$, $p = .011$, $CI = [0.06, 0.40]$). In contrast to human learners, the IO was more accurate in identifying the structure of cyclic structures (98%±8%) compared to acyclic (93%±11%) structures ($\beta = 0.45$, $t = 3.16$, $p = .010$, $CI = [0.17, 0.73]$), showing the reverse pattern to human learners. [2] This suggests that, in principle, the cyclic devices produced more information about their underlying causal structure than the acyclic devices. IO accuracy in general does not predict human trial-by-trial accuracy ($\beta = 0.05$, $t = 1.40$, $p = .178$, $CI = [-0.02, 0.12]$). As Figure 6.6 shows, the relationship between IO and participant accuracy was moderated by cyclicity with a positive association across acyclic structures ($\beta = 0.16$, $t = 3.61$,

[2]For 5% of trials in Experiment 1 and 3% in Experiment 2, participants generated such a high event density that we were not able to calculate the posterior due to there being too many possible causal paths to evaluate ($> 10^{15}$). This tended to happen if a participant intervened very rapidly, particularly on cyclic structures where each event tended to spawn many subsequent events. We simply omit these trials from current analyses.

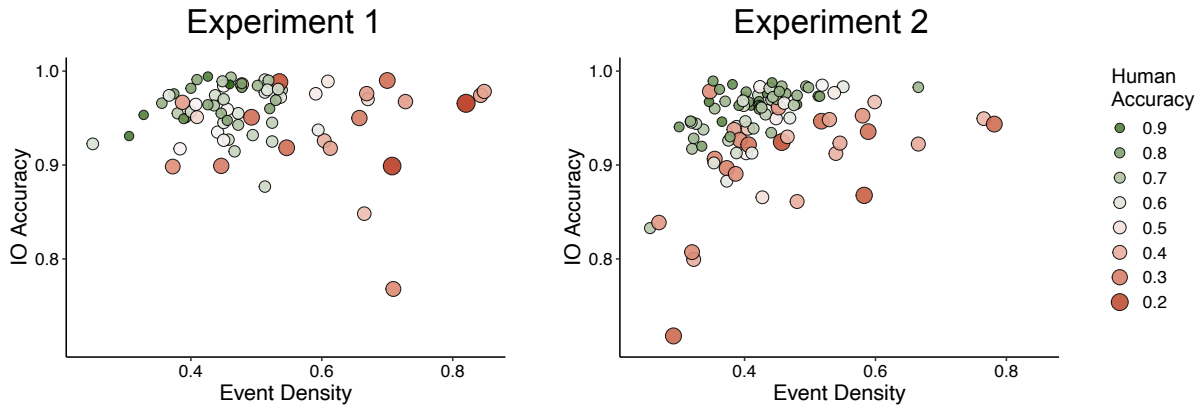


Figure 6.7: Scatterplots of average final IO accuracy (indexing evidence informativeness) and event density (indexing evidence complexity) for each participant with color and size indicating that participants' judgment accuracy. Participants with higher accuracy generated evidence that was both more informative and less complex (the upper left area).

$p = .002$, $CI = [0.07, 0.24]$) but none for cyclic structures ($\beta = -0.04$, $t = 0.84$, $p = .407$, $CI = [-0.13, 0.05]$).

We use event density (number of events per second) as a basic index for how *complex* the generated evidence was. Event density differed dramatically between acyclic (0.22±0.06) and cyclic (0.79±0.41) devices ($\beta = 1.39$, $t = 11.80$, $p < .001$, $CI = [1.16, 1.63]$) and participants' accuracy was generally negatively related to density of events ($\beta = -0.27$, $t = 3.71$, $p < .001$, $CI = [-0.41, -0.14]$). However, as shown in Figure 6.6, event density was negatively associated with accuracy on cyclic ($\beta = -0.31$, $t = 3.30$, $p = .02$, $CI = [-0.49, -0.12]$), but not acyclic ($\beta = -0.01$, $t = 0.25$, $p = .810$, $CI = [-0.11, 0.08]$) devices.

These results suggest that evidence complexity is critical in this task. For the IO, complexity is generally positively correlated with success. The more activations there are, the more information an ideal observer can use to reduce its uncertainty. However, non-ideal human learners clearly struggled to deal with complex evidence. As shown in Figure 6.7, the best-performing participants were generally those who were able to generate evidence that would have enabled the IO to be highly accurate, but that was also low in event density. We later compare computational models that model the influence of complexity on human judgments and intervention selection, capturing the qualitative differences between cyclic and acyclic cases (see the section on *Modeling the judgments*).

When to intervene We now assess whether participants' intervention choices are qualitatively consistent with the idea that they choose interventions that generate strong evidence while minimizing evidential complexity. For example, participants may choose to perform fewer interventions when experiencing a large numbers of events, as tends to occur in cyclic structures and, to a lesser extent with structures with four components (Figure 6.8). As shown in Figure 6.9,

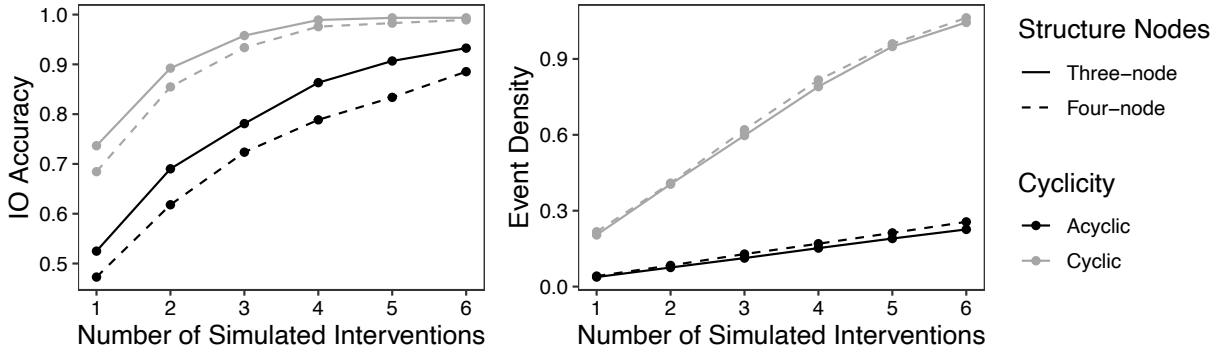


Figure 6.8: Relationship between intervention count, evidence strength (IO accuracy) and complexity (event density) in simulated causal interactions. Simulations based on randomly activating a component in a random causal system (from Figure 6.4) at $t \in \{0, 7.5, 15, 22.5, 30, 37.5\}$ seconds so a full set of six interventions would be distributed evenly across 45s.

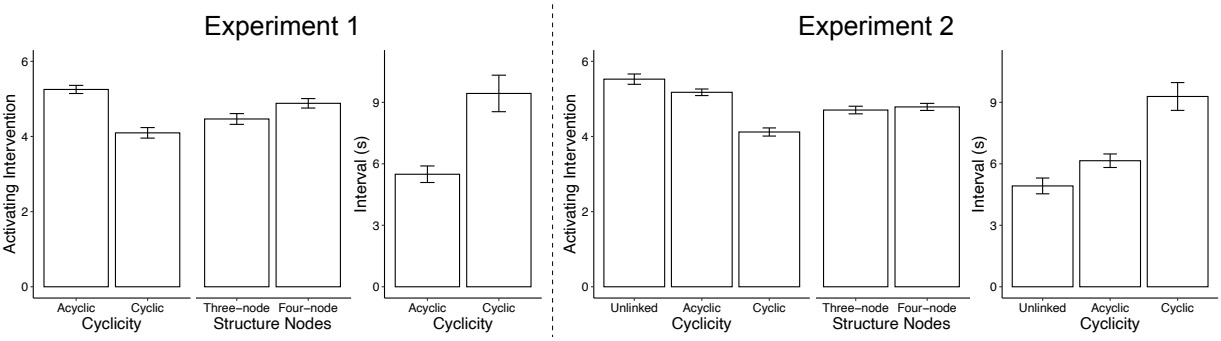


Figure 6.9: The average numbers of activating interventions used and average time intervals between activating interventions under different structures. Error bars indicate 95% confidence intervals. Each data point shows an individual's average intervals for acyclic or cyclic causal structures.

on average, participants performed 4.68 ± 1.46 out of the maximum of 6 interventions on each trial, performing about the same number in the unreliable (4.76 ± 1.41) and reliable condition (4.58 ± 1.51, $\beta = 0.12$, $t = 0.93$, $p = .36$, $CI = [-0.14, 0.39]$). However, participants performed fewer interventions on cyclic (4.10 ± 1.50) than acyclic devices (5.25 ± 1.16, $\beta = 0.79$, $t = 6.24$, $p < .001$, $CI = [0.54, 1.04]$) and fewer on three-component (4.47 ± 1.54) than four-component devices (4.88 ± 1.35, $\beta = 0.28$, $t = 2.27$, $p = .045$, $CI = [0.04, 0.53]$). These results correspond to the simulation results in Figure 6.8: In cyclic structures, even a few interventions allowed for ceiling level accuracy in principle, while the event density compounded dramatically with each additional intervention. This means that, in cyclic structures, the computational cost of additional interventions quickly outweighs the value of the new information. Event density also increased going from three- to four-component devices (at least for the structures we tested), but this increase comes alongside an increase in the amount of structure to be learned. This means

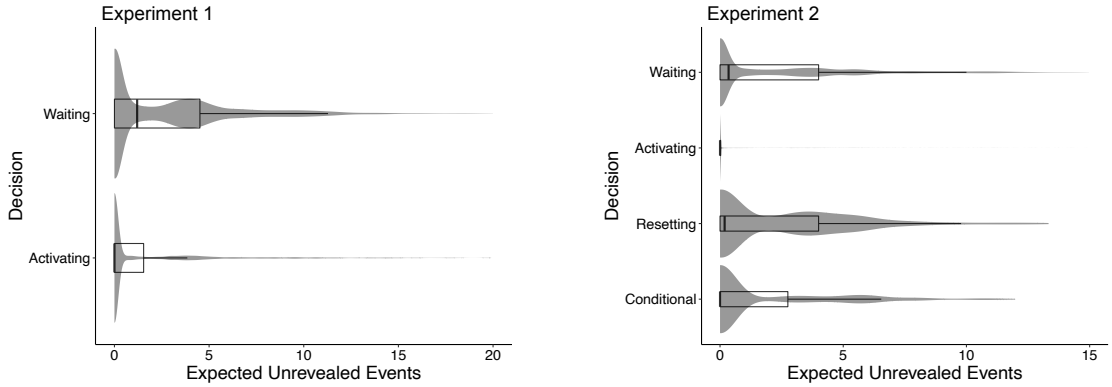


Figure 6.10: Number of expected unrevealed events across all 1-second decision windows in all trials, as a function of whether an intervention is performed in that window. Windows for which participants performed more than one intervention were excluded. The densities are scaled for each experiment to have equal maximum width. 0.6% and 1.3% of the data, from Experiment 1 and 2 respectively fall outside of the visualized area (i.e. have expected unrevealed events larger than 20).

that the evidence-strength gap between three- and four-component devices could be narrowed by intervening more frequently in four-component devices compared to three-component devices.

The average interval between each intervention depended on cyclicity ($\beta = 0.67$, $t = 7.22$, $p < .001$, $CI = [0.49, 0.86]$), with participants waiting longer before the next intervention when the structure being learned was cyclic (9.38 ± 5.94s) rather than acyclic (5.49 ± 2.64s, Figure 6.9). There was no evidence for a difference in this measure between the unreliable and reliable delay conditions (7.55 ± 5.10s vs. 7.26 ± 4.83s) or between three- and four-node problems (7.19 ± 4.92s vs. 7.63 ± 5.01s). As shown in the example in Figure 6.2b, even if the total number of events is identical, learning is easier when the events are more spread out across the trial. Intervention-spreading may be particularly important for cyclic structures where the event density is higher.

We test whether participants' tendency to wait longer under cyclic structures is driven by an anticipation of complexity. As a first pass, we calculated a moment-by-moment expectation of the level of computational cost of the upcoming evidence assuming no further intervention is performed. For this, we calculated the number of events expected to occur in the near future as a result of earlier activity (see *Modeling the interventions* for more details). We can compare the moments in which participants did nothing with those in which participants performed an intervention. As shown in Figure 6.10, people waited — i.e. did not perform any intervention — for 88% of the 1-second windows in which they could have acted (i.e. had not run out of activating interventions yet). Yet in the 12% of time windows where they *did* intervene, the number of already-expected events ($Median = 0$, $Mean = 1.86$) was lower than those where they did nothing ($Median = 1.25$, $Mean = 2.96$, Mood's median test: $\chi^2(1) = 854.35$, $p < 0.001$). This result suggests that participants tended to wait to perform their next intervention once there was not too much expected activity.

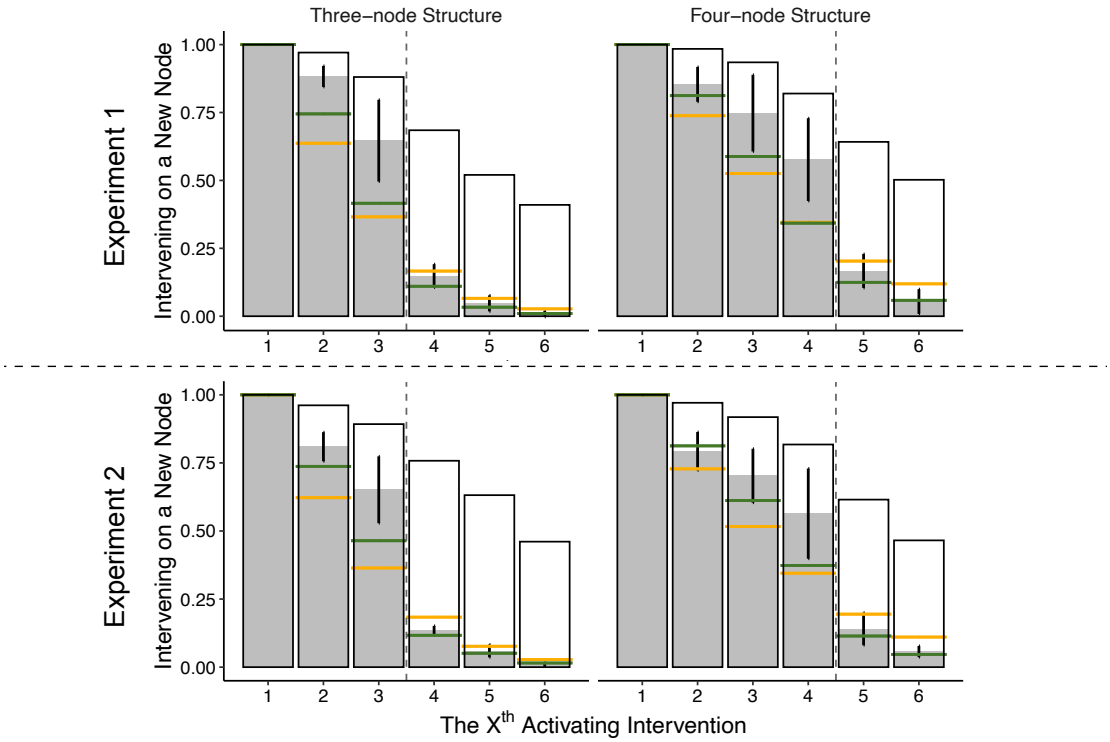


Figure 6.11: Participants' tendency to activate a node they have not intervened on previously as a function of intervention index. Black frames indicate the proportion of trials in which the participant performed at least this many interventions. Error bars indicate 95% confidence intervals. Yellow lines indicate performance under random selection $(N_{node}-1)^{(X-1)}/(N_{node})^{(X-1)}$ where N_{node} represents the number of nodes in the system. Green lines indicate the level based on the idealized information maximizing intervener who made choices at the same moments as participants and conditional on the same prior evidence. Vertical dashed lines indicate the boundary after which the learner has performed enough interventions to have tried every component once.

Where to intervene An efficient sequence of interventions involves both a healthy dose of early exploration — trying each component to learn its effects — but also an exploitative reactive focus — meaning a later tendency to repeat activating components that showed promise in producing effects. This repetition allows a learner to gather evidence about the order and the delay with which effects propagate through the system. This information is crucal for distinguishing between devices with overlapping causal structure. We see clear qualitative evidence of such exploration and exploitation in participants' choices. Figure 6.4's node shading shows the aggregate proportion of interventions on each node in each structure. Participants' interventions were relatively evenly distributed across components for most devices. They had a slight tendency to activate more causally "central" nodes (i.e. nodes that have many descendant edges; Coenen et al., 2015) in devices that have these such as on A in the two common cause structures (`Acyclic3` and `Acyclic6`).

A marker of early exploration is a tendency to initially sample components to test *without* replacement. That is, choosing something different to activate on one's second test than one's

first, and so on. Figure 6.11 shows how frequently participants selected a novel component to intervene on as a function of serial intervention position within the trial. This is shown in Figure 6.11, and we compared participants' performance against chance (i.e. the random intervener) as well as against the choices of an idealized expected-information-gain maximizing intervener (i.e. the EIG intervener, see the section on *Modeling the interventions* for more details) taking actions at the same moments as participants and conditional on the same prior evidence. For both three- and four-node structures, participants were more likely than chance to intervene on untested components until the number of interventions exceeded the number of components in the system ($ts(73) > 10.91, p < .001$). This shows that participants were not intervening randomly and suggests that they typically began by exploring system components they had not activated yet. The simulated informationally efficient intervener shows a similar pattern the first several interventions (Figure 6.11). The efficient intervener's decisions, along with those of participants become reactive to the past evidence in complex ways that do not submit to a straightforward aggregate measure. As such we will examine these choices closely through modeling in a section after the experiments.

6.2.3 Discussion

In Experiment 1, we showed that people are able to infer causal structure through active intervention in a challenging continuous-time learning setting. We found that participants, had different error patterns to an ideal observer model, in particular making more accurate judgments about acyclic structures than cyclic structures while the ideal observer had the reverse pattern. We also found that differences in accuracy across conditions were associated with differences in the character of the evidence. The informativeness of evidence predicted participants' performance in acyclic structures, but the complexity of evidence appeared to dominate participants' performance in cyclic structures where it was generally higher. Participants who were able to generate evidence that was both informative but not overly complex tended to perform best overall. We take this to support our central idea that managing computational cost plays an important role in interventional decisions and success in the real-time causal learning setting.

Intervention choices were partly shaped by a drive to control computational demands. In terms of when to intervene, participants performed fewer interventions and waited longer between them on cyclic structures that tended to produce more events. They also tended to perform more interventions on four-node structures yielding a similar number of events as for three-node structures but presumably responding to the greater initial uncertainty (larger space of structure possibilities). When the expected upcoming evidential complexity was already high, participants were more likely to wait rather than activate another component to produce more events.

In terms of which components participants would target, we found they used their interventions to systematically explore the devices, tending to select a hitherto untested component for their first

few interventions, qualitatively in line with the behavior of an efficient information maximizing agent. Participants also showed a tendency to repeat-intervene on causally "central" components once these were discovered. Note that the role of a root component activation differs in this setting to the atemporal settings studied in the past literature. Interventions on known-to-be causally central components has previously been framed as a heuristic Positive Testing Strategy (Coenen et al., 2015; Steyvers et al., 2003) on the grounds that it is often correlated with expected information yet much easier to calculate. Positive testing can be very poor in the atemporal setting because multiple causal influences from the root component overshadow one another since all effects are revealed at once. However, in the continuous-time point-event case, intervening on a suspected root will often generate rich and diagnostic evidence through the delays and order variability in the propagation of the activity through the system(Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018). We will test the extent to which participants' specific where-to-intervene choices reflect an information gain norm in our model fitting to follow (see the section on *Modeling the interventions*).

6.3 Experiment 2: Activating and blocking

Experiment 2 aims to replicate and extend the results of Experiment 1. This time, participants were not only able to activate components but they could also choose to block components, temporarily preventing them from activating until unblocked again. Intuitively, blocking permits the learner a greater degree of control over interactions with and observations of the system, as they can now isolate components to focus on, and also take control of ongoing activity in the system. On the other hand, a larger action space increases the complexity of the intervention decision-making problem. We will examine whether the relationship between ideal observer accuracy, human accuracy and event density is similar to the activation-only setting, and explore how participants use the blocking function. In particular, we will assess whether participants spontaneously use blocks to reduce the complexity of evidence without substantially reducing its diagnosticity about the causal relationships.

6.3.1 Methods

Participants

95 participants (54 female, 40 male, 1 nonbinary, aged 36 ± 12) were recruited from Prolific Academic and were randomly assigned to reliable-delay ($N = 48$) or unreliable-delay ($N = 47$) condition. They received a basic payment of £1 and a bonus depending on performance as in Experiment 1. Fourteen additional participants were tested but removed from the analysis because they reported for all the trials that the structure was completely unconnected, which was the initial default ($N = 7$), or did not perform any interventions in at least one trial ($N = 7$).

Design & Procedure

The interface was similar to Experiment 1 with a few changes. In addition to activating components, participants were also able to block components by right clicking on them and to unblock them again with an additional right click. Blocked components were marked visually by turning gray and by showing a stop sign on them (Figure 6.3b). Blocked components did not activate when they otherwise would have been caused to do so by another event or by a left click activation intervention. While participants were limited to 6 activations (as in Experiment 1), we did not limit how many times components could be blocked and unblocked.

We also extended the test set of causal devices. We added two densely connected acyclic structures (`Acyclic7` and `Acyclic8`, Figure 6.12), two densely connected cyclic structures (`Cyclic7` and `Cyclic8`), as well as two devices with no causal connections between the components (i.e. `Unlinked1` and `Unlinked2`). The inclusion of unconnected structures served to explore how people intervene under one extreme setting where no effects are ever experienced. While unconnected devices are technically acyclic, they are also qualitatively unique and as such, we treated them as a separate device type in our analyses. The new densely connected structures, by comparison, might produce particularly complex evidence and would so be particularly amenable to the use of blocks. Given the larger set of test stimuli, we did not not include a practice trial in Experiment 2. In other respects, the instructions, incentive structure, and randomization procedure were identical to that of Experiment 1.

We improved the interface in two ways. First, we were concerned that occasionally two activations of a component would overlap making them hard to distinguish visually. In Experiment 1, each activation caused the component to turn yellow for 200 ms, if two activations overlapped this would result in the component appearing yellow for longer but without a clear declination between events. In Experiment 2, we had each component turn yellow and then fade back to gray over 200 ms, this made it easier to detect distinct activation events even if their onset times were very close together. Second, to make providing judgments more seamless, participants were not required to click a "confirm" button to register when they had finished making a change to their structure judgment as they had had to in Experiment 1. One second after they stopped clicking on the edges, the state of their currently-marked structure was automatically registered as their latest judgment.

6.3.2 Results

As in Experiment 1, we first look at judgment accuracy and error patterns, and then at intervention strategies. We first focus on use of activations and then explore when and where participants use the novel blocking function.

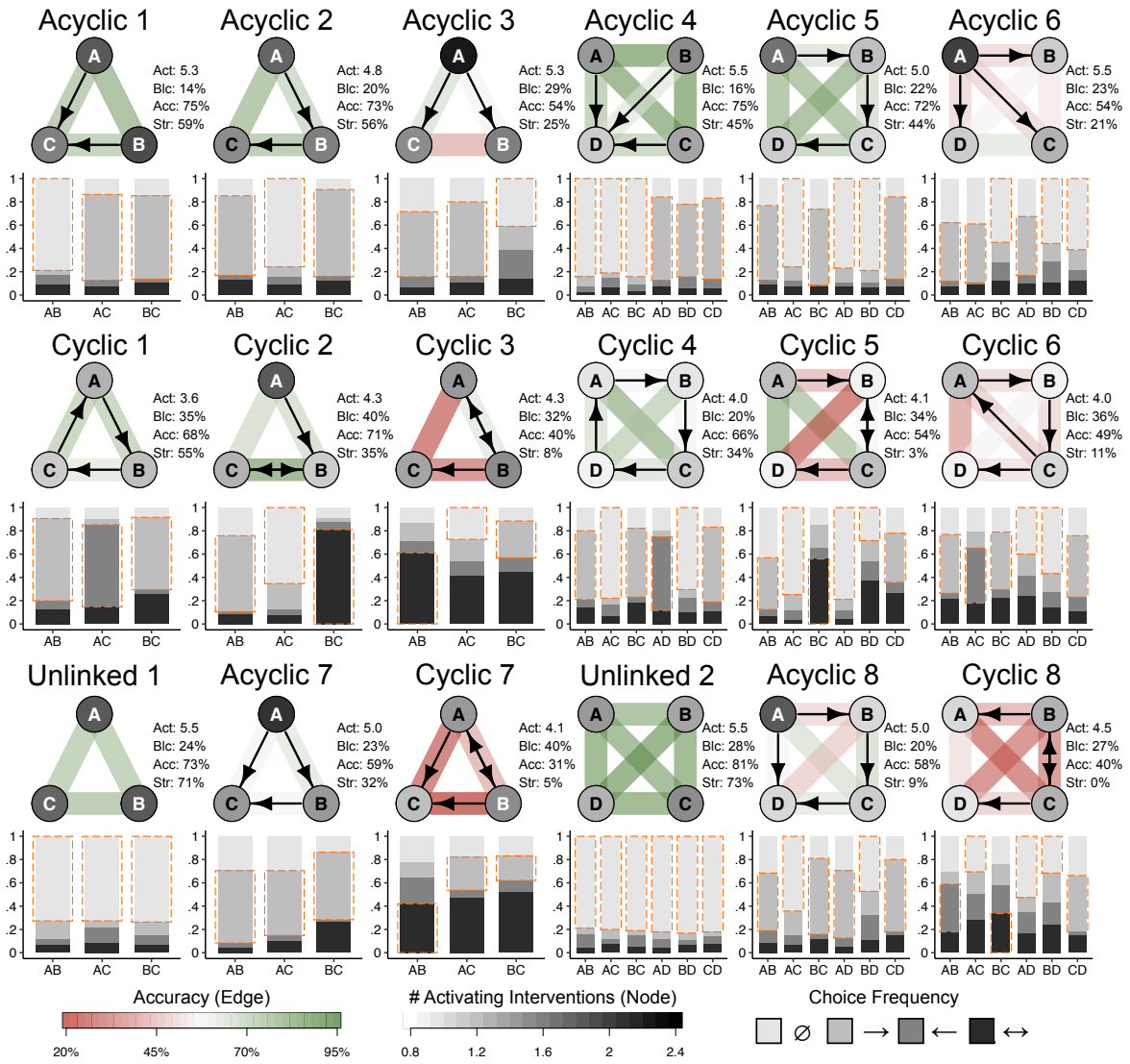


Figure 6.12: Causal link identification and activating intervention choices in Experiment 2. Color edge shading indicates accuracy. Node shading indicates activating intervention choice prevalence by component. Bar plots show the proportion of different choices on each link (e.g. for AB, "$\varnothing$" means "no connection between A and B"; "$\rightarrow$" means "$A \rightarrow B$"; "$\leftarrow$" means "$A \leftarrow B$"; "$\leftrightarrow$" means "$A \leftrightarrow B$") with orange used to highlight the ground truth. Note: Act = average number of activating interventions performed; Blc = proportion of participants who used blocking; Acc = mean accuracy; Str = proportion of participants who detected the whole structure correctly.

Accuracy Participants registered judgments 4.39 ± 2.43 times per trial. Within trials for which the answer was registered more than once (86% of all trials), final judgments were more accurate than initial judgments with $60\% \pm 34\%$ compared to $48\% \pm 24\%$ of connections correctly identified, $\beta = 0.41$, $t = 14.09$, $p < .001$, $CI = [0.35, 0.47]$. Participants' judgments became more accurate as they approached the end of the trial ($\beta = 0.11$, $t = 10.92$, $p < .001$, $CI = [0.09, 0.12]$). As in Experiment 1, we focus on the final answers as our primary measure of task performance.

Table 6.1 shows participants' accuracy separated by conditions. Performance in both reliability conditions was significantly above chance (random: 25%, reliable: $t(47) = 11.58, p < .001$, Cohen's $d = 1.67$; unreliable: $t(46) = 13.63$, $p < .001$, Cohen's $d = 1.99$). The average accuracy for all 18 structures were above chance in the reliable condition ($ts(47) > 4.01, ps < .001$) and the unreliable condition ($ts(46) > 3.39, ps < .01$) with the exception of `Cyclic7` (reliable: $t(47) = 2.01$, $p = .05$; unreliable: $t(46) = 0.56$, $p = .58$, Figure 6.12).

Unlike in Experiment 1, participants' performance only differed between unlinked and cyclic structures ($\beta = 0.72$, $t = 2.26$, $p = .04$, $CI = [0.10, 1.35]$, Figure 6.5), with a marginally significant difference between acyclic and cyclic structures ($\beta = 0.37$, $t = 1.85$, $p = .09$, $CI = [-0.02, 0.75]$). There was no main effect of delay reliability ($\beta = 0.03$, $t = 0.20$, $p = .84$, $CI = [-0.24, 0.30]$) or the number of components ($\beta = 0.08$, $t = 0.34$, $p = .74$, $CI = [-0.37, 0.53]$), nor were there any two- or three-way interactions ($ps > .10$).

Error patterns Figure 6.12 shows the types of errors people made in inferring causal structures. For chain structures (`Acyclic2`, `Acyclic5`), there were no systematic errors in mistaking them as fully-connected structures, or fork structures (less than 5%). Similar to Experiment 1, 15% of participants mistook the fork structure `Acyclic3` as a chain structure ($A \rightarrow B \rightarrow C$ or $A \rightarrow C \rightarrow B$), while 12% mistook it for a fully-connected structure by adding a directed link between two child nodes. In the case of the fully-connected structure `Acyclic7`, 15% of participants disregarded the link between $A \rightarrow C$, while 10% confused it for a cyclic structure $A \rightarrow B \leftrightarrow C$. For `Acyclic8`, 13% of participants confused $B \rightarrow C$ with $A \rightarrow C$, or $C \rightarrow D$ as $B \rightarrow D$. The error patterns did not significantly differ between unreliable and reliable delay conditions (χ^2 tests, $ps > .10$). Similar to Experiment 1, these error pattern results seem consistent with the idea that reliance on local computation and simple event order played a role in some participants' judgments (Burns & McCormack, 2009; McCormack et al., 2016; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018).

For cyclic structures, participants in Experiment 2 also performed poorly for structures that contained one or more output components. The output components were frequently judged to be constituents of the feedback loop. Participants tended to connect components whose activations often occurred close in time. This pattern was replicated in the two new structures `Cyclic7`, `Cyclic8`. For instance, participants frequently marked erroneous bidirectional $A \leftrightarrow C$ and $B \leftrightarrow C$ connections in `Cyclic7` and $A \leftrightarrow C$ and $B \leftrightarrow D$ in `Cyclic8`.

As with Experiment 1, participants could generally tell whether a structure was cyclic or acyclic regardless of whether they got all causal connections correct. They choose the correct class 70%±46% of time for acyclic structures (excluding the unlinked structures) and 80%±40% of time for cyclic structures. Participants more often mistook acyclic structures for cyclic than the reverse ($t(94) = 2.46$, $p = 0.02$).

Informativeness and event density Based on observing participants' interventions, the ideal observer performed significantly better at identifying the structure of cyclic (97% ± 9%) than unlinked (89% ± 11%) devices ($\beta = 0.74$, $t = 3.59$, $p = .004$, $CI = [0.33, 1.14]$, Figure 6.5), but not acyclic (93% ± 12%) devices ($\beta = 0.37$, $t = 1.85$, $p = .09$, $CI = [-0.02, 0.75]$). It also performed better for three-node (96% ± 10%) than four-node (93% ± 11%) devices ($\beta = 0.40$, $t = 2.62$, $p = .02$, $CI = [0.10, 0.69]$). Unlike in Experiment 1, IO accuracy positively predicted participant judgment accuracy overall ($\beta = 0.07$, $t = 3.11$, $p = .002$, $CI = [0.02, 0.11]$). Similarly to Experiment 1, the degree of correlation differed depending on cyclicity ($\beta = 0.10$, $t = 3.28$, $p = .003$, $CI = [0.04, 0.16]$), with a positive relationship in acyclic devices ($\beta = 0.15$, $t = 3.27$, $p = .008$, $CI = [0.06, 0.24]$) but no relationship for cyclic devices ($\beta = 0.04$, $t = 0.12$, $p = .22$, $CI = [-0.03, 0.11]$, Figure 6.6).

Event density differed between acyclic (0.23±0.08) and cyclic (0.71±0.39) trials ($\beta = 1.31$, $t = 11.86$, $p < .001$, $CI = [1.09, 1.58]$), and between unlinked (0.12±0.02) and cyclic trials ($\beta = 1.61$, $t = 9.71$, $p < .001$, $CI = [1.28, 1.93]$). Participants' accuracy was generally negatively associated with event density ($\beta = -0.20$, $t = -6.15$, $p < .001$, $CI = [-0.26, -0.14]$). However, this also depended on cyclicity. Event density was negatively correlated with human accuracy in cyclic trials ($\beta = -0.21$, $t = 3.88$, $p = .006$, $CI = [-0.32, -0.11]$), but there was no significant relationship for acyclic trials ($\beta = 0.10$, $t = 1.73$, $p = .12$, $CI = [-0.01, 0.22]$, Figure 6.6). As shown in Figure 6.7, participants who generated evidence with both high informativeness and low complexity performed better in general.

When to activate Participants performed 4.75 ± 1.46 activations on average on each trial. This did not differ across reliable (4.75 ± 1.42) and unreliable (4.74 ± 1.49) conditions or across three-node (4.70 ± 1.51) and four-node (4.79 ± 1.40) structures. However, participants performed fewer activations on cyclic (4.12 ± 1.52) compared to acyclic devices (5.18 ± 1.23, $\beta = 0.72$, $t = 6.97$, $p < .001$, $CI = [0.52, 0.93]$), or compared to unlinked devices (5.53 ± 0.96, $\beta = 0.97$, $t = 6.16$, $p < .001$, $CI = [0.66, 1.27]$, Figure 6.9). Participants also waited longer to perform activations in cyclic than acyclic structures (9.26 ± 5.57s vs. 6.15 ± 2.54s, $\beta = 0.53$, $t = 7.63$, $p < .001$, $CI = [0.39, 0.66]$) or unlinked structures (4.91 ± 2.33s, $\beta = 0.79$, $t = 7.57$, $p < .001$, $CI = [0.58, 0.99]$), and longer in acyclic than unlinked structures ($\beta = 0.26$, $t = 2.74$, $p = .02$, $CI = [0.07, 0.44]$, Figure 6.9). There were, again, more expected unrevealed events on seconds where participants waited (Median= 0.44, Mean= 2.58) than those where they chose to activate a component (Median= 0, Mean= 1.29, Mood's median test: $\chi^2(1) = 2423.60$, $p < 0.001$, Figure 6.10).

When to block Participants performed blocks on 27% of trials (949 times in 1710 trials, 0.55 ± 1.17 per trial). 75 of 95 participants (79%) used blocking at least once. Given that the frequency of blocking was much sparser than activations, we coded trials as 1 (used blocks) or

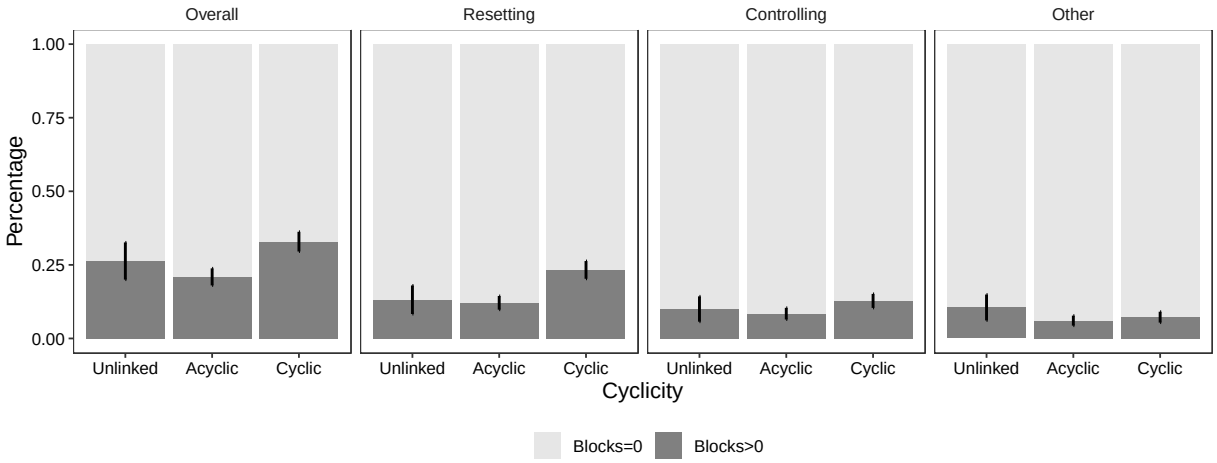


Figure 6.13: Percentages of blocking behaviors. Error bars indicate 95% confidence intervals.

0 (no blocks) in our statistical analyses and fitted logistic regression models to explore when blocking was used.

We found that the propensity to block differed neither between the reliable (28%) and unreliable (26%) delay conditions ($\beta = 0.12$, $z = 0.87$, $p = .386$, $CI = [-0.15, 0.38]$), nor between three-node (29%) and four-node (25%) devices ($\beta = 0.07$, $z = 0.50$, $p = .620$, $CI = [-0.20, 0.33]$). However, participants were more likely to use blocks in cyclic (33%) than acyclic (21%) structures ($\beta = 0.61$, $z = 5.16$, $p < .001$, $CI = [0.37, 0.84]$). Surprisingly, propensity to use blocks when facing unlinked (26%) structures did not differ significantly from cyclic or acyclic structures ($ps > .05$). We had anticipated participants would be less likely to block in unlinked structures since a key function of blocks in this setting is to manage evidential complexity, and in the unlinked structures this is always minimal. We speculated that some uses of blocks might be spurious since we did not limit their use. For example, sometimes participants may have blocked and unblocked components simply to kill time until the end of the trial, especially after they have used up the activating chances and the system has been silent for a while. To further explore how participants used blocking, we focused on categorizing blocking actions, focusing on two plausible goals of blocking that have distinct empirical signatures: (1) Blocking in combination with activating to control for confounding causal paths and (2) blocking to reset the device.

For both of these uses, we derived simple operationalizations. We take "Controlling" blocks to be those that appear to be used as a way to perform a controlled test, essentially isolating a sub-network made up of all the components except the blocked one(s). This way, sub-networks can be investigated through an activation without the possibility of interference from any pathways through the blocked component. In contrast, the "Resetting" category includes those where the learner blocks and then unblocks a component before performing another activation, without activating any other component in the interval while the component is blocked. In the current

setting, Resetting blocks serve to short-circuit ongoing chains of causal effects of previous interventions, essentially resetting the mechanism so that subsequent tests can be performed without interference. Both forms of blocking reduce density of events experienced during the trial but do so in conceptually different ways (See examples in Figure 6.2b). "Resetting" blocks — where the next action is to unblock the same component — made up 55%, "Controlling blocks" — those followed by an activation of a different component — made up 24%, and the remaining 21% were classified as "Other". This nuisance category includes cases where a block is performed by a participant who has no activations remaining or performs no subsequent activation or unblocking action before the end of the trial.

Participants performed more Resetting blocks in cyclic (23%) than acyclic (12%, $\beta = 0.79$, $z = 5.53$, $p < .001$, $CI = [0.51, 1.07]$) or unlinked (13%, $\beta = 0.69$, $z = 2.96$, $p = .003$, $CI = [0.25, 1.17]$) devices. There was no difference between acyclic and unlinked structures ($\beta = 0.09$, $z = 0.38$, $p = .704$, $CI = [-0.41, 0.56]$). Similarly, participants performed more Controlling blocks in cyclic (13%) than acyclic (8%, $\beta = 0.47$, $z = 2.72$, $p = .007$, $CI = [0.13, 0.82]$) devices, but unlinked structures were not significantly different than either (10%). Participants performed slightly more Other-type blocks in unlinked structures (11%) than acyclic structures (6%, $\beta = 0.64$, $z = 2.24$, $p = .025$, $CI = [0.06, 1.19]$) but not cyclic structures (7%). In sum, both Resetting and Controlling were used more on cyclic devices.

We checked whether participants used Resetting and Controlling blocks in ways that make sense from a bounded rationality perspective. We assume that Resetting is useful at moments where expected future complexity is high, while Controlling blocks should be used when people expected low complexity (i.e. when they were ready for the next activating intervention). In line with this, the number of expected unrevealed events was higher for seconds in which Resetting blocks were performed ($Median = 0.21$, $Mean = 2.62$) than those where Controlling blocks were performed ($Median = 0$, $Mean = 2.12$, Moodâ€™s median test: $\chi^2(1) = 14.69$, $p < 0.001$, Figure 6.10).

We asked whether performing blocks is related to participants' accuracy. We compared accuracy of different structures in Experiment 1 and 2 (Figure 6.4 vs. Figure 6.12) and found that the accuracy of full loops `Cyclic1` ($t(167) = 3.62$, $p < .001$) and `Cyclic4` ($t(167) = 2.53$, $p = .012$) significantly differed between two experiments. We checked whether the blocking function accounted for the difference. For `Cyclic1`, participants in Experiment 2 performed at least one Resetting (26%) were more accurate than those who did not make this kind of block ($t(93) = 2.42$, $p = .018$, Cohen's $d = 0.56$). However, Controlling blocks (17%) did not make a significant difference ($t(93) = 1.48$, $p = .141$). This finding was replicated in `Cyclic4` that accuracy was positively associated with the use of Resetting blocks (16%, $t(93) = 3.58$, $p < .001$, Cohen's $d = 1.01$), but not with Controlling blocks (6%, $t(93) = 1.87$, $p = .06$). This indicated that Resetting blocks may be more helpful than Controlling blocks. Given that the performance was better for several cyclic structures, there was some benefit to having the blocking ability.

Where to activate Similar to Experiment 1, participants tended to explore the devices initially by activating untested components. Participants were more likely than chance to activate an untested component with their second and third activating interventions (and their fourth for four node devices, $ps < .001$), which is in line with how the informationally efficient intervener behaves (Figure 6.11).

Discussion In Experiment 2, we allowed participants to use blocking as a tool for causal learning. The addition of blocking made the action space larger but also gave participants more fine-grained control over the learning input, allowing them to not just inject excitation into the system but also to selectively inhibit it. As found in Experiment 1, evidence informativeness positively predicted participants' performance in acyclic structures while evidenctial complexity negatively predicted performance in cyclic structures. Accuracy was less strongly associated with structure cyclicity than in Experiment 1, which may in part be due to the fact that blocking helped participants to accommodate and counteract the differences in excitability and ambiguity characteristic of interactions with the different causal devices. We also replicated the finding that people performed fewer activations and waited longer to perform their next activation in cyclic structures where expected computational cost was generally higher.

Participants used blocking in only a quarter of trials. However, when blocking was employed it was used in sensible ways that primarily managed inferential complexity. Participants blocked more often in cyclic than acyclic devices and did so when many events could be expected to occur in the near future. This is consistent with the assumption that learners take management of computational cost into consideration when choosing how to intervene in real time. We categorized blocks according to two potential goals: Resetting the system and Controlled testing — combining a block with an activation to test a subsystem in isolation. Both of these strategies were more likely to employed after moments of high expected complexity.

6.4 Modeling the judgments

The following two sections detail our quantitative analysis of the role of complexity in shaping participants' causal judgments and intervention choices. We compare a set of computational models to demonstrate that: (1) Participants' causal judgments were affected by evidential informativeness and complexity and (2) participants' interventions strike a balance between the informativeness and complexity of future evidence. Readers less interested in technical detail can safely skip ahead to the General Discussion.

To quantitatively test the idea that evidence complexity is not just positively related to informativeness, but also impacts human performance directly, we built a computational-level model that assumes human causal judgments $q \in \mathbf{Q}\{X \rightarrow Y, X \leftarrow Y, X \leftrightarrow Y, X \varnothing Y\}$ are a noisy version of the ideal observer's posterior marginalized across connections IO_q, where the noise degree

Table 6.2: Judgment model fits.

	CV	BIC	τ_1	τ_2	Best
Experiment 1					
Random	-5215	10430			9 (16)
IO	-4151	8259		0.62	28 (26)
IO/N	**-4045**	**8057**	0.78	0.32	37 (32)
Experiment 2					
Random	-10310	20620			13 (16)
IO	-8030	15983		0.59	26 (2)
IO/N	**-7894**	**15644**	0.67	0.34	56 (77)

Note: The "best" column displayed the number of individuals best-fit by each model according to CV, with BIC results in the brackets.

depends on the density, and hence complexity, of the evidence. We capture this with a dynamic softmax function (Luce, 1959):

$$P(\text{judgment} = q) = \frac{\exp\left(\text{IO}_q / (\tau_1 N + \tau_2)\right)}{\sum_{q' \in \mathbf{Q}} \exp\left(\text{IO}_{q'} / (\tau_1 N + \tau_2)\right)} \tag{6.1}$$

where N denotes a trial's event density (average number of events per second). The judgment temperature component is thus a linear function of events $f(N) = \tau_1 N + \tau_2$ with two parameters $\tau_1, \tau_2 \in (0, +\infty)$ that are constant across trials, while N varies across trials depending on what interventions are performed and how the system reacts to them.

We fit this model with participants' choices across two experiments and compare it to a baseline model that made random judgments, and a informativeness-based model that only considers IO judgments by omitting $\tau_1 N$ from Equation 6.1. We used hold-one-device-out cross-validated log-likelihood as our primary measure of model fit but also include BIC for completeness and comparison with past work (Tauber et al., 2017). Our cross-validation scheme is conservative, since it requires a unified explanation for human data despite different causal devices exhibiting markedly different characteristic dynamics.

Table 6.2 shows the results. In both Experiment 1 and 2, the model that combines informativeness and complexity outperforms the informativeness-only model. Individual results are relatively similar between the two models in Experiment 1, but more-strongly favor the combination model in Experiment 2 where there were more data points for a single person (12 vs. 18 data points in Experiment 1 vs. Experiment 2). These results suggest that, while more complex evidence carries more information on average, its complexity takes a toll on human performance, presumably due to our cognitive limitations.

6.5 Modeling the interventions

The final part of our analysis describes a computational account of complexity-sensitive intervention selection and compares it to participants' intervention choices in both experiments. Intervention selection is the problem of choosing what to do now, in order to support future learning. Normatively, this depends on the learner's prior over causal structures $P(S)$ at the point of the decision, which in turn depends on the already-observed data $\mathbf{d}_t$ and earlier interventions $\mathbf{i}_t$. In our first experiment, where participants can activate but not block, each participant must choose, at each moment in time, between intervening on one of the components or doing nothing, leading to the intervention space $\mathbf{I} = \{a_A, a_B, a_C, \varnothing\}$ for three node systems. If the learner has used all their activation chances, this reduces to just the option of doing nothing $\varnothing$. In our second experiment, where participants and models could also block components, the action space is larger, including actions that toggle the block status of each node such that it becomes blocked if currently unblocked or unblocked if currently blocked (e.g. $\mathbf{I} = \{a_A, a_B, a_C, b_A, b_B, b_C, \varnothing\}$).

While in principle this intervention decision needs to be made constantly, at every instant throughout the trial. In practice we simplified our analyses by assuming that learners make exactly one intervention selection decision per second.[3]

6.5.1 Expected information gain

Information gain (IG) is a common currency for measuring the value of evidence for an ideal learner (Shannon, 1948; Nelson, 2005; Coenen, Nelson, & Gureckis, 2019). The goal is to select the intervention (or sequence of interventions) that is *expected* to have high information gain or, in other words, that best reduces the learner's uncertainty. To do this exactly, one must quantify how much every possible intervention decision i_t^* is expected to reduce future uncertainty about the structure of the causal system given the current beliefs and marginalizing over consideration of possible future evidence. We take a greedy approach in favoring actions expected to maximally reduce future uncertainty at this point but without considering potential subsequent actions.[4] The learner's uncertainty at time t_x can be measured by calculating the Shannon entropy $H(S)_{t_x}$ of the current prior $P(S)_{t_x}$ based on all the evidence experienced so far:

$$H(S)_{t_x} = \sum_{s \in S} P(s)_{t_x} \log_2 \frac{1}{P(s)_{t_x}} \tag{6.2}$$

[3]Thus, we do not attempt to predict when, within a specific 1-second window, any action would be taken but just what action, if any, is performed in each window. Occasionally participants performed more than one action within a 1-second window. This was very rare though, occurring in only 0.41% of windows in Experiment 1 and 0.36% in Experiment 2. For simplicity, we simply treated these multi-action windows as missing data and modeled the other >99% of trials.

[4]This is a common choice due to submodularity results about the diminishing utility of planning ahead in active learning problems (Guillory, 2012).

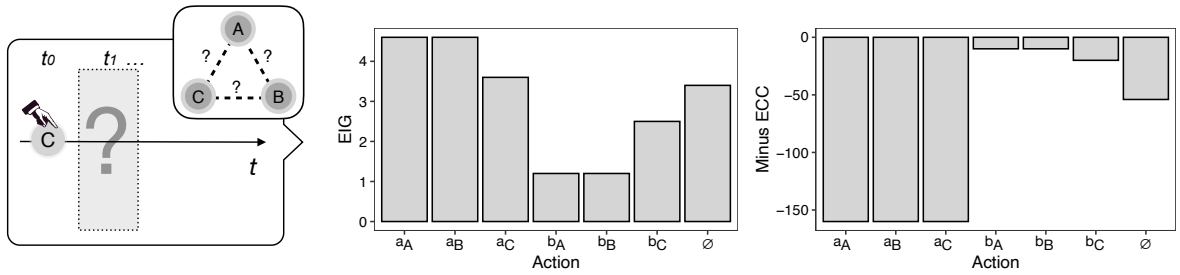


Figure 6.14: Example of expected information gain (EIG) and expected computational cost (ECC). The learner activated C at t_0 and is now deciding what to do at t_1. The notions of a_X, b_X, and $\varnothing$ stand for choices to activate X, block X, or do nothing, respectively. Both EIG and ECC are temporally discounted. ECC was calculated based on expected local events with a polynomial function.

The ideal calculation of future information should consider all possible future evidence $\mathbf{o}$ up to some future time point t_y, given the hypothetical action i^*. However, unlike the atemporal setting, the outcome space here is continuous, meaning we must approximate this integral by sampling a subset of possible futures. We achieve this by simulating a set of possible outcome sequences $\tilde{\mathbf{o}}$ under different structures. We further assume the number of samples simulated under each structure is based on the structure's (current) prior probability (Nelson, 2005). [5] For each simulated future $\tilde{o} \in \tilde{\mathbf{o}}$ we compute the information gain as:

$$\mathrm{IG}(i^*, \tilde{o})_{t_x}^{t_y} = H(S)_{t_x} - H(S, i^*, \tilde{o})_{t_y} \tag{6.3}$$

and expected information gain as:

$$\mathrm{EIG}[i^*]_{t_x}^{t_y} = \sum_{\tilde{o} \in \tilde{\mathbf{o}}} \mathrm{IG}(i^*, \tilde{o})_{t_x}^{t_y}. \tag{6.4}$$

Note that in this setting, the anticipated information results not only from the focal action choice i^*, but also from other recent actions and effects that may still be expected to produce further effects and evidence. This means that one can often expect substantial information to be forthcoming even when choosing not to act ($i^* = \varnothing$).

Figure 6.14 shows an example where the learner has already intervened, activating C at $t = 0$. Even though no effects have occurred yet, they are considering what to do one second later, at $t = 1$. The value of doing nothing ($\varnothing$) is relatively low at this point as it only includes expected information resulting from the previous intervention. The learner expects less value from activating C a second time than for activating something else, since they expect to learn about the consequences of C from their first intervention. The blocking actions ($\{b_A, ..., b_C\}$) also

[5] All results are based on sampling 512 event sequences from each window. We selected this number for computational practicability and it is a multiple of 64 which is the cardinality of three-component structures and square root of the cardinality of four-node structures. We checked that this sample size resulted in stable and consistent results by replicating the simulation process with different seeds.

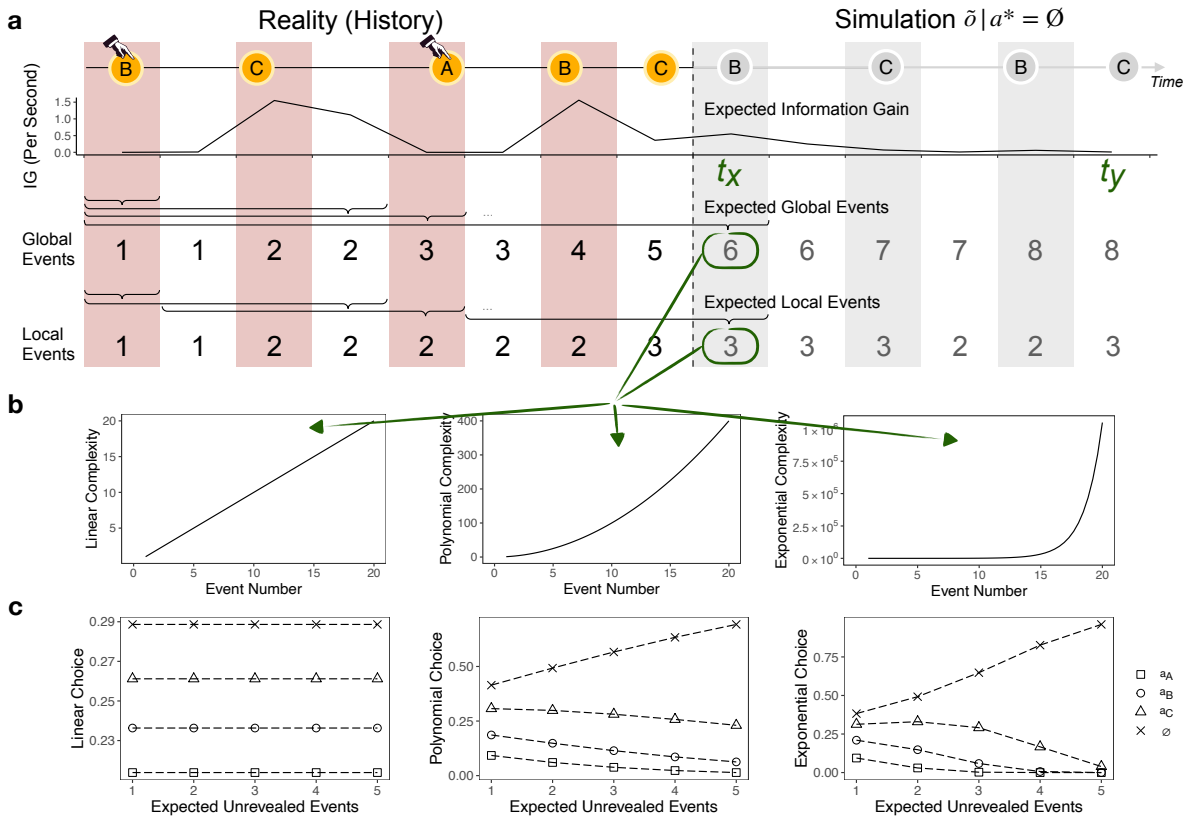


Figure 6.15: An illustrative example of historical and *expected* Information Gain (IG), alongside historical and expected global events and local events. a) Pink portion (left of t_x) = Window-by-window IG about true structure; Global events (since start of observation); and Local events (within the last 4 seconds). Gray portion (right of t_x) Expected upcoming information, global and local events. b) Three possible computational cost functions of event number. c) A demo of how different complexity functions react to the number expected unrevealed events under the softmax function. Expected evidence was generated from a $A \to B \to C$ chain.

have low expected information since, at this stage, they would only serve to block potentially informative dynamics produced by the previous intervention.

6.5.2 Expected cost of inference

There are various ways to measure the computational costs of integrating causal structure evidence. Our inference framework works by considering the various pathways connecting the interventions and effects under each considered structure. The number of paths scales rapidly with the number of plausibly-related effects (Figure 6.2a), meaning a naïve realization of our ideal observer performs an amount of computation that scales super-exponentially in the total number of events observed so far. Nevertheless, considering all past events back to the beginning of time, which we call the *global event* set, is clearly infeasible outside of very simplified toy settings. Inevitably, practical constraints come into play such as excluding from consideration events that occurred

long enough ago to have a negligible chance of having caused the most recent effect. For simplicity, in our primary analyses we simply assume learners are focused on a 4-second "backtracking window" (see Figure 6.15a; c.f. Gerstenberg et al., 2013). That is, we assume learners enumerate and consider causal pathways involving events or interventions from up to several seconds prior to the moment at which the inference is taking place. We chose 4 seconds as the window size as this is long enough to include all plausible causes for any newly occurring event under our delay regime. We refer to these as the *local event* set and assume the learner reasons over a rolling window of local events throughout the trial. This results in a measure of inferential cost that shifts throughout a trial as a function of the number of recent events (see Figure 6.15a). We will also examine other choices of the window size in Table B.1.

While idealized Bayesian inference also requires estimation of the evidence in parallel under all possible hypotheses, in practice it is implausible that a bounded learner would consider the entire hypothesis space at the same time since this quickly becomes intractable as the number of components increases. For instance, there are 4,064 possible structures linking 4 components together and this would increase to 1,048,576 if there were 5 components in the system. A recent proposal for how learners mitigate the complexity of structure inference in the natural world is that they consider hypotheses sequentially. For example, in the atemporal dataset setting, it has been argued that participants consider evidence under a single favored hypothesis at a time, regenerating or adapting this hypothesis only to the extent that it fails to explain the most recent evidence (Bramley, Dayan, et al., 2017; Bonawitz et al., 2014).

Since humans must, by necessity, find a more scalable approach to causal inference than our normative algorithm in order to to succeed in the wild, we think of the idealized Bayesian inference as an upper bound on the computational cost of inference. We explore intervention behavior under several plausible inference-complexity-scaling functions based on either the global or local number of events n and some base parameter c. These functions, including linear $O(n)$, polynomial $O(n^c)$, exponential $O(c^n)$ scaling, differ at how fast the cost grows with the increase of event numbers (Figure 6.15b).

Similar to expected information gain, we can also anticipate computational cost (CC) of integrating future evidence. This involves counting the events occurring in simulated outcomes $\tilde{o} \in \tilde{\mathbf{o}}$. For each hypothetical future time point considered, we count the recent events $n(t)$ and compute the consequent complexity of performing inference about how these events could relate:

$$\mathrm{CC}(i^*, \tilde{o})_{t_x}^{t_y} = f_{\text{complexity}}(n_{t_x}^{t_y}, c). \tag{6.5}$$

where we will later allow $f_{\text{complexity}}$ to be of linear, polynomial or exponential form with some parameter c, in either the anticipated local or global events (see Figure 6.15).

We can then compute the *Expected Computational Cost* (ECC) by summing over $\tilde{o} \in \tilde{\mathbf{o}}$

$$\mathrm{ECC}[i^*]_{t_x}^{t_y} = \sum_{\tilde{o} \in \tilde{\mathbf{o}}} \mathrm{CC}(i^*, \tilde{o})_{t_x}^{t_y} \tag{6.6}$$

6.5.3 Resource-rational intervention utility

According to the resource-rational framework (Lieder & Griffiths, 2020), the expected utility of an action $\mathbb{E}[U(i^*)]$ to a bounded learner balances expected reward and cost of computation. In our case, this results in the following equation:

$$\mathbb{E}[U(i^*)] = \sum_{t=t_x}^{t_y-1} R(t) \cdot \left[\mathrm{EIG}[i^*]_t^{t+1} - \omega \cdot \mathrm{ECC}[i^*]_t^{t+1} \right] \tag{6.7}$$

where we assume a 1 second granularity for measurement, and where ω scales the cost component to align it with the epistemic reward scale of bits, the sum aggregates the expected future gains and costs over future seconds up until t_y, with $R(t)$ as a discount function which diminishes the utility of information and the dis-utility of computational costs the further into the future they occur. In our case this is simply done according to how long the trial remains to end (i.e. chance to affect the bonus):

$$R(t) = 1 - \frac{t}{t_{end}} \tag{6.8}$$

The ideal t_y horizon should be the end of the learning episode t_{end} (i.e. 45 s in our experiments), but we found no substantial impact upon our choice predictions beyond $t_x + 6$.[6] Finally, a resource rational learner should behave according to:

$$\underset{i^* \in \mathbf{I}}{\mathrm{argmax}} \left[\mathbb{E}[U(i^*)] \right] \tag{6.9}$$

Figure 6.15 visualizes the various elements of a trial that combine into our resource rational algorithm and Figure 6.14 shows an example in which information gain and inferential complexity differ in the choices they favor and hence trade off. In sum, our framework captures how a resource rational agent should decide when and where to intervene to support their causal structure learning. We will compare human interventions against the predictions of this modeling framework.

[6] Intuitively, this horizon is reasonable here for several reasons: (1) The rational temporal discount factor makes the distant future less important. (2) Expected information gain under the "greedy" assumption of no future activations approaches zero after a handful of seconds, by which time even the most complex causal systems have had enough time to loop through all their causal relationships at least once. (3) The inherently stochastic delays combined with the complicated causal interactions and compounded by the learner's uncertainty thereof leads to complicated simulated dynamics whose predictive power rapidly drops toward chance beyond a few seconds (cf. Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018).

Table 6.3: Intervention model fits.

	CV	BIC	τ	ω	θ	Best
Experiment 1						
Baseline	-17033	34027	0.31			7 (14)
EIG	-16588	33109	3.01		10.80	10 (7)
EIG-$\text{ECC}_{\text{Global}}$	-16456	32880	2.67	4.89×10^{-3}	8.99	11 (4)
EIG-$\text{ECC}_{\text{Local}}$	**-16415**	**32792**	2.27	8.82×10^{-3}	7.91	46 (49)
Experiment 2						
Baseline	-43042	85997	0.27			2 (3)
EIG	-40507	80887	1.98		8.13	19 (51)
EIG-$\text{ECC}_{\text{Global}}$	-40507	80898	1.98	1.69×10^{-6}	8.13	14 (2)
EIG-$\text{ECC}_{\text{Local}}$	**-40462**	**80804**	1.86	1.59×10^{-3}	7.63	60 (39)

Note: The "best" column displayed the number of individuals best-fit by each model according to CV, with BIC results in the brackets. BIC for the fully random baseline was 97210 in Experiment 1 and 262515 in Experiment 2. Parameters reported were based on BIC results.

6.5.4 Model fitting

Our primary class of models is based on the utility function specified in Equation 6.7. That is, one that is sensitive to both expected information and computational cost. The measurement of this cost is formed as (linear, polynomial, or exponential) complexity functions of (global or local) real-time expected event numbers. The complexity function controls how quickly the model expects computational costs to increase as the event number increases. For intervention decisions, this affects how sensitive the model predicts people will be in favoring choices less likely to result in large numbers of events in the future. Note that only polynomial or exponential scaling can capture the phenomenon that the more the expected unrevealed events caused by previous interventions, the greater the likelihood of avoiding future activating interventions (Figure 6.15c). We used polynomial scaling with a generic exponent of 2 as the primary form here while the results of other complexity functions with different exponents or bases can be found in Table B.2.

To investigate whether participants are sensitive to both expected information gain and expected computational cost, we also examined a purely information-driven model that removes ECC from Equation 6.10. However, a model that considers computational costs could easily beat such a purely information-driven model given the fact that greedy-EIG underestimates the value of waiting when opportunities to intervene are finite.[7] In contrast, the vast majority of time windows in the human data did not contain an action.[8] Therefore, we included a constant bias against action that increased the probability of not acting: $B(i) = 1$ if $i = \varnothing$ and $B(i) = 0$

[7]Conceptually, this is because it does not incorporate the expected utility of saving an action for later use.

[8]This is also owing to the fact that time spent on activities that were not directly connected to the learning process, such as marking their answers and moving the mouse, was taken into account.

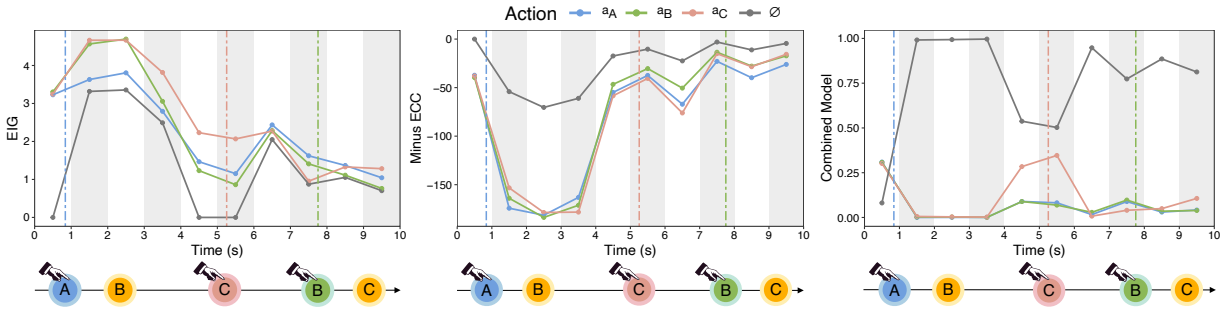


Figure 6.16: Example of real-time model prediction for a participant in reliable condition of Experiment 1 facing $A \rightarrow B \rightarrow C$ structure. Lines and points show instantaneous value for each potential intervention (colours) or non-intervention (black). Dashed vertical lines show participants interventions. Model takes earlier interventions and observations into consideration and predicts value of intervention choices for each 1-second window (marked by vertical white/gray shading). Parameters of the combined model based on EIG + local polynomial cost model fit to this individual. Model fit is the product of likelihoods of the chosen action or non-action in each window.

otherwise. This allows for a fairer comparison between dynamic-cost-dependent and cost-free models. If our EIG-ECC model outperforms the EIG model, this means that participants timed their interventions in a reactive way to cope with the expectation of computational cost rather than simply avoiding action with a constant probability across time.

We assumed stochasticity in participants' intervention choices captured by a softmax function (Luce, 1959) over the resultant values. The resource-rational prediction is:

$$P(\text{intervention} = i) = \frac{\exp\left((\mathbb{E}[U(i)] + \theta B)/\tau\right)}{\sum_{i' \in \mathbf{I}} \exp\left((\mathbb{E}[U(i')] + \theta B)/\tau\right)} \tag{6.10}$$

The model fitting procedure is similar to what we used for the judgment models. We provide hold-one-device-out cross-validation results and BIC results at both the aggregate and individual levels. As shown in Table 6.3, across both experiments, models that considered both expected information gain and inference cost outperformed pure information driven models. The best variant for both experiments was one that anticipated costs on the basis of a polynomial function of the expected local events. Models including both information and costs also better fit more individuals in both experiments than the other models we considered (78% of participants were fit best by one of the cost-dependent models, 63% people were fit by the local cost model specifically). Figure 6.16 gives an example from Experiment 1 in which the combination of expected information and cost give a better account of participants' intervention choices than either does alone. In Appendix B.2, we used the fitted parameters to simulate interventions and judgments and showed that these align qualitatively with the human patterns.

The fact that the boundedly rational models outperformed pure information seeking models corroborates our central idea that participants' interventions were shaped both by how much information they expected to gain and by how hard they would have to work to process potential future information. Note that while we did not present them for space reasons, model variants sensitive to only cost but not information, perform worse than all the models we present irrespective of how the cost is calculated. These models invariably favor waiting or blocking over activating components.

More individuals in Experiment 2 were best fit by the cost-free model according to BIC. This suggests that cost-free and cost-dependent models did not differ as much as in Experiment 1 when explaining human interventions. This might be due to the fact that the computational cost component of the model predicted learners should block fairly frequently, while participants blocked less often than predicted in general. We suspect that this is partly due to a preference for simplicity but in terms of strategy choice rather than evidence, with blocking strategies being intuitively more involved. Furthermore, our models so far only consider information gain and computational cost of the current intervention, while as discussed, people are likely to plan ahead when using blocking, for instance combining a block with a subsequent activation, which goes beyond the capability of this greedy model.

6.5.5 Prospective vs. retrospective complexity

We explored whether expected computational cost — which depended on both recent events but also how many events are anticipated to happen in the near future — can be substituted with a simpler retrospective computational cost consideration — based only on how many events have occurred recently. To test this, we ran retrospective variants of each model in Table B.3 finding that these were always a slightly worse fit than their prospective versions. This could be because, while the retrospective approach captures a sensible and simple heuristic of waiting until the system is quiet, this behavior can also be accounted for by expected complexity. Moreover, retrospective complexity is insensitive to earlier learning about the structure within a trial. For instance, one might have learned that the current system is highly excitatory (captured by the evolving prior shaping expected complexity) but that activity might have died down by the time of the next intervention.

Retrospective complexity also only accounts for when to intervene but not where. In contrast, expected complexity serves as a guide for both when and where to intervene. For example, if a learner has already discovered a component that seems to generate a large number of events, they may decide not to activate it again. To test whether cost-dependent choice of where to intervene is a significant feature of our participants' intervention selections, we also fit resource-rational models to only the time windows where participants made activations. At the aggregate level, the EIG-ECC model (cross-validated log-likelihood: -4788 in Experiment 1 and -9054 in

Experiment 2) only had minor advantages over a pure EIG model in predicting these windows (cross-validated log-likelihood: -4789 in Experiment 1 and -9054 in Experiment 2). However, it did also better capture 31% of individuals. The individuals best-fit by the EIG-ECC model in terms of the windows in which they acted had better performance than 55% people who were best-fit by an EIG only model (accuracy: 67% vs. 59%, $t(144) = 2.22$, $p = .028$, Cohen's $d = 0.38$), and better performance than the 14% people best-fit as selecting components to intervene on randomly (accuracy: 67% vs. 49%, $t(74) = 3.97$, $p < .001$, Cohen's $d = 0.99$).

6.6 General discussion

In a dynamically unfolding world, uncovering causal relationships requires online control and processing of continuous sensory information. To learn about how the world works, one must choose *where* to act, *how* to act, and *when* to do so while also tracking what happens before, during, and after one's actions. In this paper, we investigated human learning in a setting where learners use freely timed interventions to investigate the underlying causal structure responsible for devices' patterns of real-time component activations. We investigated what factors affected the quality of their inferences, and what strategies they used to choose and time their interventions. We hypothesized that computational limitations, and a rational anticipation thereof, would play a key role in shaping real time active learning. Thus, we endeavored to quantify the actual and anticipated computational cost of the evidence stream in our task and used model fitting to show that this could help explain both human judgments and intervention patterns.

6.6.1 What we found

Our empirical findings fall into two classes: (1) Insights about features of real time causal systems that determine how easily people can learn them and (2) insights about how people choose interventions to support their learning.

What experimental factors affected participants' learning? Across both our experiments, participants had more success identifying acyclic than cyclic structures while an ideal observer model showed the reverse pattern, highlighting a fundamental divergence between ideal and bounded learning. The ideal observer benefits from the higher density of events produced by feedback loops, essentially because it is able to enumerate and marginalize over the many possible causal explanations for the data, and make ideal use of the rich timing information. Meanwhile, participants' ability to do this was presumably limited by their information processing capacity, leading to a kind of "less is more" phenomenon (cf. Gigerenzer & Todd, 1999) in which simpler evidence was often more valuable to them even when less normatively informative. In line with this, we showed that human accuracy patterns can be accounted for through a corruption of the ideal

observer that assumed bandwidth limitations on the processing of evidence, such that inferential noise and probability of error increase with the compounding effect of event density, potentially more than counteracting the value of the additional information.

While past work has demonstrated that people are sensitive to delay reliability, and use delay information in addition to order to shape causal structure judgments (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Greville & Buehner, 2010), reliability had little impact on performance here. Delay condition did not make a statistically significant difference to accuracy in either experiment suggesting reliable delays may be less critical in the active learning setting, where interventions can be dynamically adjusted to accommodate experienced variability. Another possibility is that our delay manipulation was too subtle. However, our unreliable condition had a sevenfold wider standard deviation which is both salient in visualizations of the trials and commensurate with past work (Bramley, Mayrhofer, et al., 2017). We note also that the top 10% of performers were almost all in the reliable condition (100% in Experiment 1 and 90% in Experiment 2). We take this to suggest that delay reliability is important for achieving high accuracy. Additionally, reliable condition participants were better at identifying that full-loop cyclic structures than unreliable condition participants which might suggest that reliable delays allowed extended temporal patterns like periodicity of cyclic activations to contribute to structure identification.

We found several other systematic judgment errors. Some participants mistook chain structures as fully-connected, marking extraneous indirect links from the root components to distal child components. This lines up with the findings of a number of atemporal causal learning studies (Fernbach & Sloman, 2009; Lagnado & Sloman, 2004) as well as studies that have used continuous valued variables (Davis et al., 2020). This pattern may reflect the "local computations" idea that people often focus on subparts of the larger system, such as on pairs of variables, and so experience an appearance of direct causation when observing indirectly connected components. Participants were also quite likely to mistake fork structures as chain structures. Since the outputs of a fork would invariably occur in some staggered pattern, this seems straightforwardly consistent with occasional fallback on a heuristic of taking temporal order to directly reflect causal order (cf. McCormack et al., 2016; Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018).

Among the cyclic structures, participants frequently mistook output components as constituent elements of their upstream feedback loops (e.g. in `Cyclic3`, `Cyclic5` and `Cyclic6`). This is not unreasonable because these output components tended to activate almost in lockstep with components of the feedback loop that produced them. While informative to an ideal observer, the subtle differences in inter-event delays between components that formed part of the loop, and components that formed the loop's outputs were presumably difficult to process for bounded human learners.

How do people choose what interventions to perform? Participants generally performed fewer activations in cyclic structures. Multiple activating interventions in cyclic structures could quickly compound the complexity of the subsequent evidence, presumably overwhelming bounded learners and motivating learners to avoid this. For four-node structures, more activations were required to achieve an equivalent level of certainty as in three-node structures, and participants performed more activations on these problems especially in Experiment 1. More generally, participants appeared to adjust their interventions depending on the current and anticipated event density. They tended to perform more activating interventions at moments when the number of expected unrevealed events was low, and were more likely to wait or use blocks to reset the system when the number of expected unrevealed events was high. Strikingly, in both experiments, the participants who performed well were those who managed to generate evidence that was relatively more informative and less complex (Figure 6.7).

In Experiment 2, participants did not use blocks nearly as often as activations, but most participants did use them and did so particularly in excitable structures and at moments of high anticipated complexity. Participants performed better on the full-loop structures when allowed to block, compared to Experiment 1 (`Cyclic1` and `Cyclic4`). Full-loop structures are intuitively far easier to understand when exhibiting a single oscillating cycle of activity. If participants performed a second activation intervention before such a cycle died out, they would face confusing evidence patterns with two excitations traveling around the system in tandem, potentially overtaking one another and going in and out of sync. Resetting blocks in particular seemed to help learners control such complicated scenarios.

6.6.2 Resource-rational active structure learning

The bounded nature of cognitive computation has long been discussed in relation to models of human learning (Anderson, 1990; Simon, 1982). While early research conceptualized the role of cognitive resources qualitatively, more recent studies have aimed to quantify cognitive costs and estimate boundedly rational norms (Griffiths et al., 2015; Lieder & Griffiths, 2020; Vul et al., 2014). Utility functions that combine both expected rewards and computational costs have been shown to better capture a variety of human behaviors including estimation (anchoring-and-adjustment, Dasgupta et al., 2017; Lieder, Griffiths, Huys, & Goodman, 2018), planning (Callaway et al., 2022), information sampling (Petitet et al., 2021), decision making (Gershman, 2020), and communication (Hawkins et al., 2021). The current paper extends this line of research to the problem of real-time active causal learning. By building and comparing computational models, we firstly showed that participants' causal judgments depended on both the informativeness and the complexity of the generated evidence. More importantly, we then showed that in addition to the standardly-considered exogenous costs of interventions (Coenen et al., 2015; Coenen, Ruggeri, et al., 2019), people also care about the internal costs that arise from integrating different forms

of causal interaction data. That is, learners were sensitive to the fact that information following an intervention has to be processable to be useful.

Specifically, out of the measures of complexity we examined, a polynomial function of inference-relevant events (those in the recent past and expected in the near future) best captured the influence of complexity on intervention choice, and we found that prospective complexity as well as retrospective complexity contributed to participants' choices. While it would be premature to take this functional form as final, or to make a judgment about whether participants under- or over-anticipated the actual effect of complexity on their inferences, we feel this reflects an intuitively sensible and plausible sensitivity to local events capturing the fact that inferential complexity compounds as the number of actual causal relata increase (Van Rooij et al., 2019; Bramley, Dayan, et al., 2017; Fernbach & Sloman, 2009).

The issue of just how complexity scales raises a question as to what inference process learners actually used in this task. Although many papers, including this one, have laid out computational-level mechanisms of causal structure induction (Rottman & Hastie, 2014; Pearl, 2000; Griffiths & Tenenbaum, 2009), these are typically intractable, requiring a run time that scales often far worse than exponentially in the number of relata. As has been argued forcefully elsewhere (Van Rooij et al., 2019), this makes most computational-level models "non-starters" as process accounts of human inference in natural settings, since any plausible account will have to deal with more than a handful of components or events without requiring a time to compute that is beyond the lifespan of the organism (or even the universe). We note that the resource-rational framework adds another layer of computation, which is itself intractable. We use it here to establish that people are sensitive to information and computational cost but we do not provide a recipe for how learners anticipate these costs, given that this depends critically on their inferential processing. Human learning is necessarily more piecemeal and approximate and indeed, human responses are much noisier than our ideal observers'. There are some promising avenues for process accounts that can model aspects of this variability and noise. Simulation-based (Gerstenberg et al., 2021), summary statistics (Gong & Bramley, 2020, 2023a), and incremental search (Bramley, Dayan, et al., 2017) algorithms have all been proposed in recent years as aspects of how learners simplify and approximate solutions to structure learning.

When considering complexity, it is perhaps surprising that there was not more difference in performance between 3- and 4-variable problems since the latter involve an order of magnitude more hypotheses. However, this is in line with recent incrementalist accounts. It has been argued that learners form one or a few hypotheses at a time (Bonawitz et al., 2014), or focus on subparts of the larger system (Fernbach & Sloman, 2009; Davis et al., 2020). These accounts are better able to scale up to inferences among more relata (Bramley, Dayan, et al., 2017). Bramley, Dayan, et al. (2017) show that in inference from interventions in the atemporal covariation setting, people rely on sequential local changes to gradually update their beliefs to incorporate new evidence. Compared to maintaining a global prior, this localist approach may help people to deal with

situations involving more than a handful of variables without invoking an exponential increase in computation or catastrophic loss of performance. Thus we conclude from this pattern, that however people manage the complexity of real time causal structure inference, they do so in such a way that they are affected by the number of events, but less by the total number of components. Indeed, the run time of our our ideal observer was far more sensitive to the number of paths it had to evaluate per possible hypothesis than the number of hypotheses it evaluated.

6.6.3 Insights for a process-level model of real time active causal learning

Causal structure induction We have shown how complexity effects human judgments from a computational level perspective, essentially shaping the nature of the optimization problem faced by active causal learners in real time (Marr, 1982). Recent work on observational causal learning has generally found simple event *order* to be a strong driver of structure judgments, with delay expectations tending to have smaller and subtler effects on inferences (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Valentin et al., 2020). As such, some form of a simpler endorser heuristic (Bramley et al., 2015; Fernbach & Sloman, 2009) which assumes that people attribute an effect to the most recent event may do a reasonable job of describing some participants judgments. Such a heuristic, however, does not explain participants' good performance in identifying acyclic and cyclic classes. Taking fork structures for example (`Acyclic3` and `Acyclic6`), the order of child node activations would often reverse following repeated activations of the root node, but participants rarely reported bidirectional connections between these child components. More fundamentally, a learner that relies on a temporal order heuristic would lack the necessary representation of the hypothesis space and uncertainty needed to guide intervention selection. We would need a separate account for how people make intervention choices, and a linking model for how the inferences and activation choices are connected with one another.

One other finding that a comprehensive process account would need to accomodate, is that in cyclic structures, participants frequently drew a bidirectional link between the output component and the component in the loop that tended to activate closest in time (`Cyclic3`, `5-7`, and `8`), despite the fact that participants were trained to expect longer delays than this between truly causally related events. One interpretation is that repeated spontaneous activations are a prototypical signal of cyclic relationships (Valentin et al., 2020). Our participants may have reasoned in terms of this abstract feature, and applied it without thinking more carefully about the exact timing at which the events happened (Goodman et al., 2011). Alternatively it could be that our pre-training on delays was not sufficient to overrule a prior expectation of causal contiguity, that is, that cause and effect events happen close in time (Hume, 1740). Indeed, past studies have required strong manipulations to override this preference (Buehner & May, 2004; Buehner

& McGregor, 2006). More work is needed to better understand how temporal delays affect causal judgments.

Intervention One long-standing debate centers on the question of whether human active learning and intervention choice involves anticipation of information at all, or whether it relies on heuristics such as simple endorsement (Bramley et al., 2015), positive or confirmatory testing (Coenen et al., 2015; Steyvers et al., 2003). In the current setting, one reasonable heuristic might be to explore components until effects are discovered. If a component appears to produce multiple effects, a learner might repeat-test it, or probe the components at which the effects occured. In this way, learners might follow a kind of extended positive testing strategy in which they focus their energies on components deemed to be to produce effects so as to gather evidence "by making the machine go". This reflects the rationale behind positive testing that has featured in the literature on atemporal active causal learning (Coenen et al., 2015; Steyvers et al., 2003; Austerweil & Griffiths, 2011; Bramley, Dayan, et al., 2017; Klayman & Ha, 1989). However, distinct from the atemporal setting, there is lots that can be learned by repeatedly testing suspected root components in our experiments, meaning it is hard to distinguish whether the repeat selection of root components was driven by explicitly computing expected information gain or by following a simpler strategy such as combining random exploration with positive testing.

Additionally, it is possible that people choose when to intervene separately to where to intervene, for instance using current complexity as a way to decide when to perform one's next intervention and then selecting this without regard to anticipated complexity. To resolve these questions about psychological processing, future studies could set up continuous-time active causal learning scenarios that pit the predictions of heuristics against those of computational norms. However, the current work shows that whatever heuristic or adaptive toolbox is proposed, to fully capture intervention choice, it must include strategies that are at least be somewhat responsive to experienced and anticipated complexity.

6.6.4 Future questions

Some questions related to continuous-time causal learning remain for future research. One open question is where *cyclic structures* fit into the overall landscape of causal cognition. A full representation of a cyclic system seems to demand a temporal dimension and predictions are generally sensitive to the system's current state. In our experiments, participants performed well above chance for most connections in most cyclic structures without extra training, and were able to reliably determine if a structure contained a cycle even though they were less accurate in identifying the exact structure. This suggests that they can understand cyclic relationship relatively intuitively. Nevertheless, some cyclic structures may be particularly challenging for people to understand, and the reasons for this may go beyond what is captured by our general computation

cost account. For example, participants performed relatively poorly in identifying the internal structure of cyclic structures with an output of a loop as mentioned in the paragraph above. An analogous ambiguity in reality could arise wherever it is unclear which events are pure effects (with no potential to control the system dynamics, such as symptoms of a disease) and which are constituent parts of the system's feedback loop (such as the pathogen). If one is interested in controlling a cyclic system, it is important to identify and act on components that are inputs to, or constituent parts of the feedback loop, rather than pure effects, since only by doing so can one nudge the system toward whatever state one wants it to take. This makes it valuable to explore more factors that affect learning and control in cyclic systems.

Another open question concerns the relationship between temporal and covariation-based causal learning. One possibility is that these depend on separate learning processes, but it also seems likely that there are points of overlap. For example, people may extract covariation information from continuous-time evidence through some process of abduction and discretization. Furthermore, interventions might help to create data that is more amenable to these forms of summarization. Better understanding of human causal induction requires us to go beyond covariation-based causal Bayesian networks (Pearl, 2000), but this should not involve discarding the insights they have provided in the search for a unified account for causal learning. Our current paradigm simplifies causes and effects to point events with no measurable duration. However, actual events are often extended in time in complex ways and many require reset or refractory period between occurrences. Therefore, it might be informative to also consider a setting in which causes must be reset or take time to recover to make this paradigm more comparable to the statistical-based causal learning.

6.6.5 Conclusions

Everyday experience is rich with events that reoccur and can be causally related in ways that allow us to predict, control, and make sense of what has happened and what is likely to happen next. While previous research on active causal learning has often sidestepped the temporal dimension, in this paper we show that human learners are sensitive to time, not just in terms of how it impinges on what can be learned from evidence in principle, but also in terms of how it shapes the practicalities of gathering and interpreting that evidence. Our experiments and modeling show that participants' causal judgments depend on not just the informativeness but also the complexity of evidence they gather, and that they adapt their actions to the ongoing event dynamics during learning so as to strike a balance between expected information gain and anticipated inferential complexity. These results contribute to our understanding of causal inference in continuous time, incorporate a new dimension to the study of human active learning and offer new directions for research into human learning.

Chapter 7

Evidence from the future

> "On principle, it is quite wrong to try founding a theory on observable magnitudes alone. In reality the very opposite happens. It is the theory which decides what we can observe."
>
> *Albert Einstein*

Outcomes of any scientific experiment or intervention will naturally unfold over time. In Chapter 5 and Chapter 6, we have seen how people process observed evidence, while we know little about how people consider the "unobserved evidence" — evidence that is on its way to come. In this chapter, I investigate how people make causal inferences from measurements over time, focusing on how they may extrapolate potential future evidence based on present trends and incorporate it when making causal judgments.

Across three experiments, I had participants observe experimental and control groups over several days post-treatment in a fictional biological research setting. I 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. I 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. The

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.

Content from this chapter is a reprint of the material as it appears in Gong & Bramley (2023b).

7.1 Introduction

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 (parecetmol) 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 (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Stephan et al., 2020; Greville & Buehner, 2010; Lagnado & Sloman, 2006; Bechlivanidis et al., 2022; Gong & Bramley, 2023a; Greville et al., 2020; Buehner & May, 2003; Buehner & McGregor, 2006). 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 (Shanks et al., 1989; Greville & Buehner, 2010; Lagnado & Speekenbrink, 2010). 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 (Grice, 1948; Anderson & Sheu, 1995). This applies to not only type-level judgments that reflect beliefs about which causal events cause which type of effect events (Greville & Buehner, 2007, 2010; Buehner & May, 2003; Buehner & McGregor, 2006), but

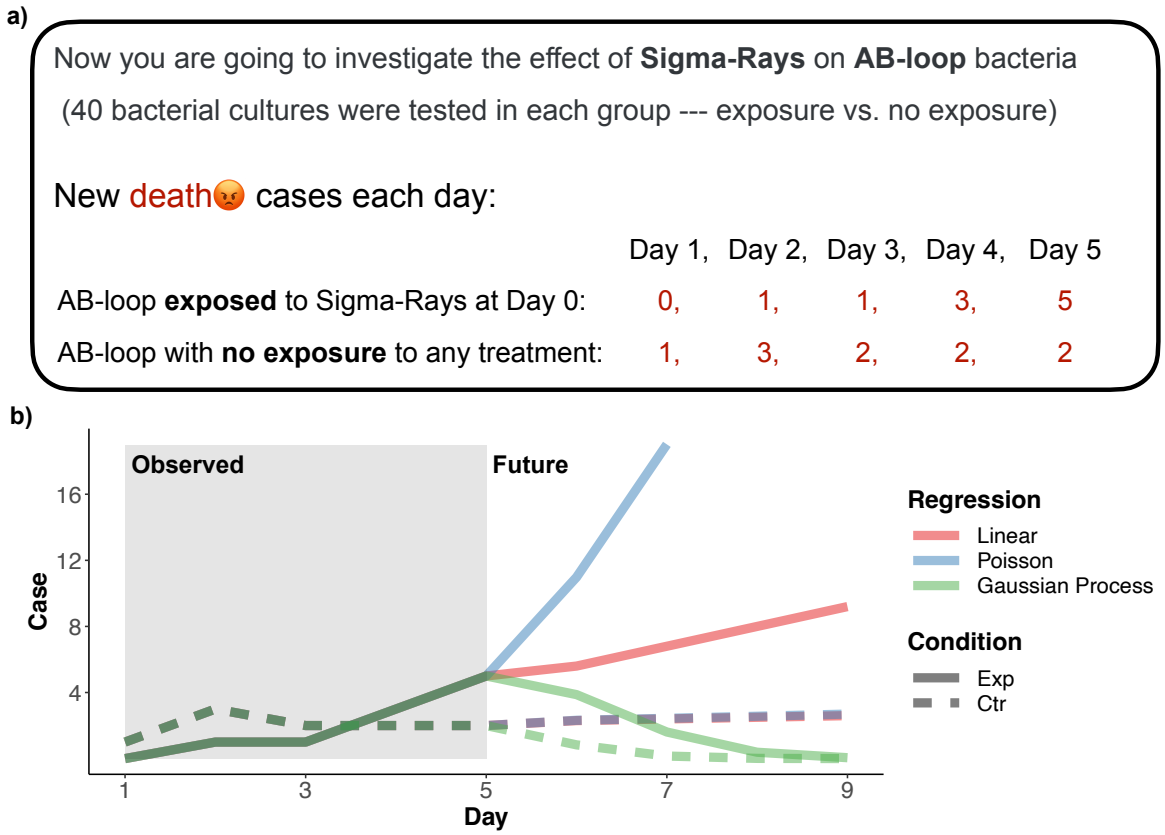


Figure 7.1: An example stimulus material of the current study (a) and the corresponding extrapolation results of how the new case will be in the future given different regression models (b). The Poisson regression would predict the experimental case as 0 at Day 9 due to the cumulative cases have exceeded the max sample size. The Gaussian process regression was based on RBF kernel (E. Schulz et al., 2017).

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 & 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 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 & 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 & 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 7.1a, 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 7.1b 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 (Shtulman & Morgan, 2017; McCoy & Ullman, 2019; Lucas & Kemp, 2015). 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 (Lucas et al., 2015; E. Schulz et al., 2017; B. Zhao et al., 2022; Hahn & Warren, 2009). We examine whether people further apply their generalizations from evidence to their causal judgments (Johnson et al., 2016).

[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 7.1: 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.

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 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.

7.2 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 & 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.

7.2.1 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. [2]

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 7.1, 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 7.1).

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 & 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 7.2.

[2]Material, data, analysis code of all experiments are available at https://osf.io/h2y3g/.

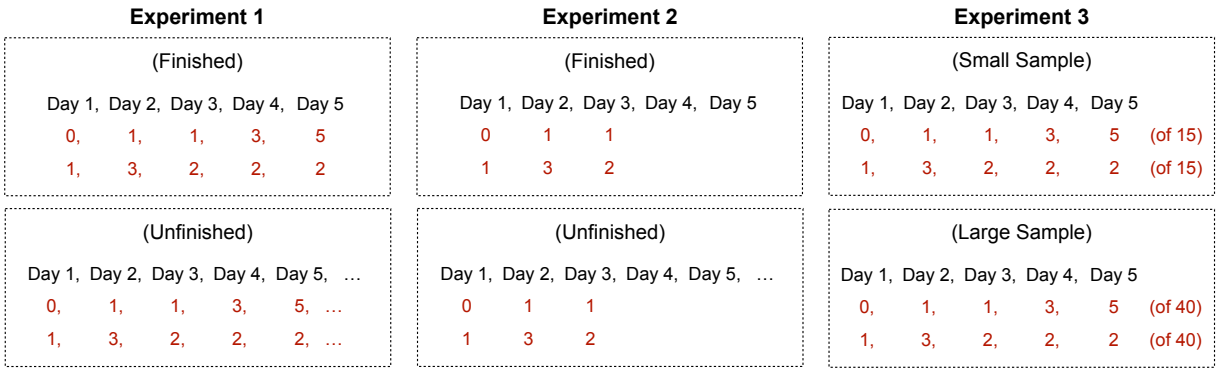


Figure 7.2: Stimuli displays under different conditions. Participants observed the number over days in a similar format shown in the Introduction with specific modifications illustrated in this figure. In Experiment 1 and 2, the sample size was disclosed to participants in text.

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 & 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).

7.2.2 Results

A three-way mixed ANOVA Analysis was performed. As shown in Figure 7.3, 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 7.4, 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 or the visual format difference (the dots in the Unfinished condition), we conducted a supplementary

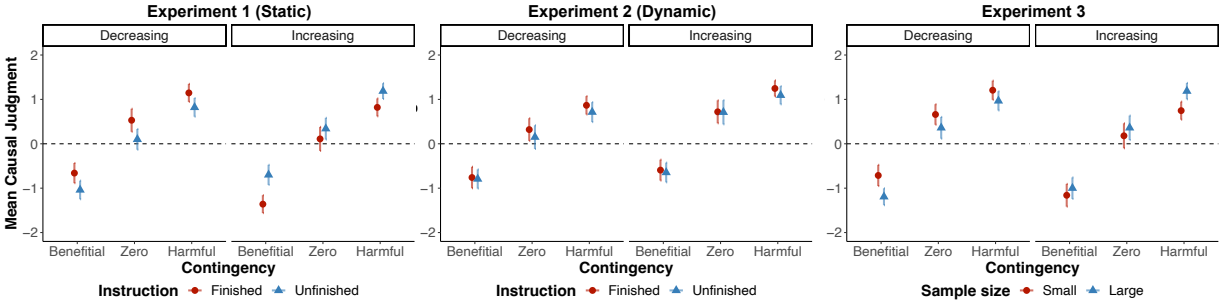


Figure 7.3: Means of causal judgments under different contingency and experimental conditions. Participants judged the influence of treatment on a scale from -3 (definitely beneficial) to 3 (definitely harmful). Dashed lines indicates the middle level when it is not sure whether the treatment was harmful or beneficial to the survival of the bacteria cultures. Error bars indicate 95% confidence intervals.

experiment (N=200) by only keeping the visual format differences between two groups (see Appendix C 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.

7.3 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.

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

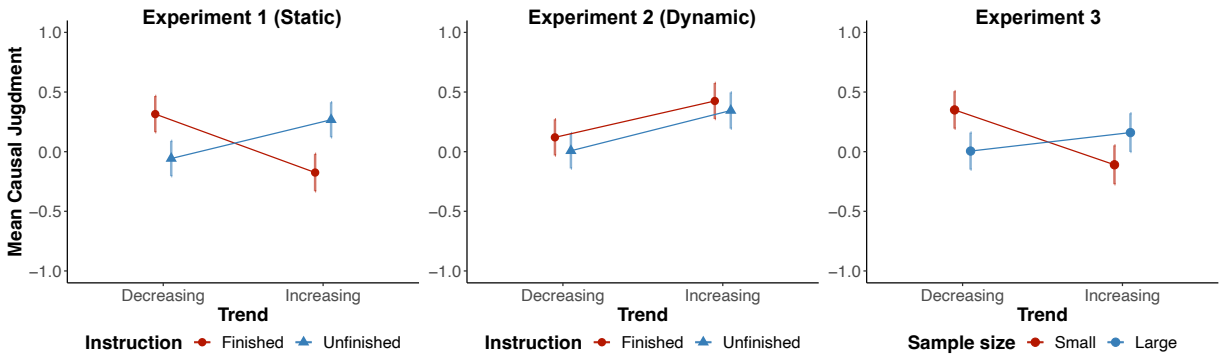


Figure 7.4: Means of causal judgments under Decreasing vs. Increasing trends across experimental conditions. Participants judged the influence of treatment on a scale from -3 (definitely beneficial) to 3 (definitely harmful). Error bars indicate 95% confidence intervals.

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 7.2). 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".

7.3.2 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.

7.4 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 by use of a dynamic display, participants

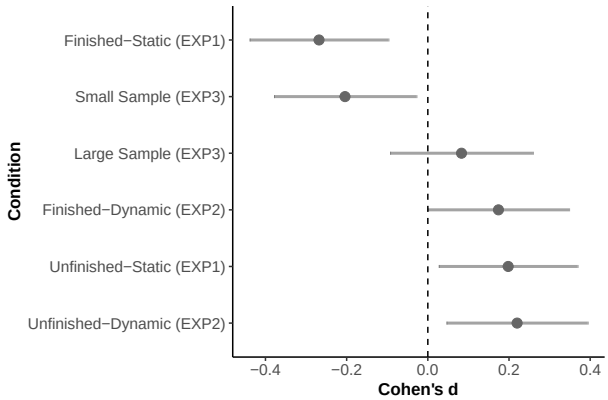


Figure 7.5: The Cohen's d effect size pooled out from Increasing-Decreasing simple effect tests in different conditions across experiments. Negative values mean participants prioritized contiguity over trend, while positive values mean participants prioritized trend over contiguity. Error bars indicate 95% confidence intervals of Cohen's d estimates.

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.

7.4.1 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 7.2). 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.

7.4.2 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 7.3). 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 7.4, 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 7.5. 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).

7.5 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 (Cheng, 1997; Buehner et al., 2003; 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 (Greville & Buehner, 2007; Pacer & Griffiths, 2012; Buehner, 2006). 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; Yu et al., 2014; Gong et al., 2023). This becomes even more critical in the kinds of dynamic contexts and complex causal effects that are ubiquitous in everyday life (Coenen, Nelson, & Gureckis, 2019; Anvari et al., 2022).

7.6 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.

Chapter 8

General discussion

Causal reasoning plays a critical role in modern scientific inquiry. Researchers have developed formal theories for experimentation and data analysis, enabling the discovery of causal relationships from empirical evidence. However, the scope of causal reasoning extends far beyond the confines of the scientists' laboratory. Its origins can be traced back to ancient times when our survival as a group hinged on our understanding of cause and effect. Today, it remains an integral part of our pursuit of a better life, permeating every moment. We learn over time, updating our beliefs every day as we encounter new experiences. We learn from time, entangling the relationships among events based on how they unfold chronologically. Learning becomes an inherent part of our lives, as we obtain knowledge from the events that affect us directly or those that affect those we care about.

In this thesis, I investigate how people learn causal structures from events unfolding in continuous time. It is surprising that despite decades of developing causal learning theories, none have effectively addressed the challenges posed by continuous-time data. Equally surprising is the fact that despite decades of empirical studies on temporal causal learning, a systematic exploration of diverse causal structures — encompassing generative vs. preventative and acyclic vs. cyclic relationships — remains absent. These intriguing gaps persist because empirical and theoretical advancements are interdependent, each helping the progress of the other (Guest & Martin, 2021). Therefore, to investigate this specific subject, this thesis embraces a computational cognitive psychology approach. It constructs new quantitative theories that describe the process of inferring causal structures through the timing of events; it also tests humans on causal learning tasks that encompass a range of causal structures. This iterative interplay between modeling and empirical data seeks to deepen our comprehension of human causal learning, particularly within the context of learning from events in continuous time.

8.1 Summary of the main findings

In Chapter 4, I develop a rational framework of temporal causal learning, and demonstrate the extent of sensitivity people exhibit towards this framework. This is an important first step, as it establishes the benchmark performance against which human behavior will be compared. The rational model developed in this chapter encompasses a wide range of tasks that have been collected from previous studies, as well as three novel tasks designed in this thesis.

While the rational framework can partially accounts for individuals' performance, human behavior may not always align with normative expectations. Temporal causal learning poses a unique "double trouble". Firstly, temporal information inherently entails greater complexity compared to atemporal information, as the observations within a sample are not independent. Multiple possibilities arise concerning how an event is generated and its potential connections to other events. Secondly, learning in a temporal context necessitates learners to simultaneously observe ongoing events and contemplate their interrelationships. A successful learner may need to simultaneously summarize past occurrences, pay attention to the present, and imagine potential future. Both of these issues might present difficulties for human intellect. Consequently, in order to better elucidate human behavior, I expand upon the rational framework by developing additional cognitively plausible models:

In Chapter 5, I investigate how people learn causal structures that involve generative and preventative relationships. People demonstrate the capability to identify the correct structure; however, they also exhibit susceptibility to the influence of intertwined evidence from different variables. It suggests that people may not engage in precise inferences for each event but rather segment the evidence and compress information. To better account for participants' choices in the task, I develop a summary-statistic model. This model incorporates three cognitive features, namely mental simulation, local evidence, and local computation, providing a more comprehensive explanation of participants' causal judgments in the task.

In Chapter 6, I investigate how people actively learn causal structures, by implementing interventions over time. Participants strategically time their interventions to ensure digestible evidence, balancing the desire for informative evidence while avoiding inferential complexity. I use the resource-rational framework to provide a comprehensive explanation of participants' intervention decisions, quantifying the informativeness and complexity of the evidence. The results indicate that participants' decisions regarding where and when to intervene align more closely with the predictions of the resource-rational model, which achieves a balance between information gain and computational cost.

One foundational question regarding temporal causal learning is how temporal information is mentally represented. Time has a distinctive feature of continuous unfolding, where each passing moment transitions the present into the past and the future into the present. This dynamic nature of time potentially influences how we perceive and process temporal information. In Chapter 7,

I show that when people observe evidence, their attention extends beyond the present moment. They imagine what may occur in the future and incorporate imagined evidence into their causal inferences. This finding may challenge the conventional understanding of temporal evidence and how it is used in human causal reasoning.

8.2 Theoretical implications

Model-based cognition Although the temporal and atemporal data may appear markedly dissimilar, this thesis demonstrates their compatibility within the Bayesian modeling framework. The rational model of temporal causal learning, as shown in Chapter 4, expands upon the Bayesian probabilistic model to effectively capture temporal information. This extension indicates that, akin to atemporal causal learning, temporal causal learning can be conceptualized as a model-based Bayesian belief updating process.

To study the mental process of temporal causal learning, I direct my focus not solely towards the literature of causal learning, but also a broader range of model-based cognition, such as physical reasoning (Battaglia et al., 2013; Ullman et al., 2018; Ludwin-Peery et al., 2021), planning (Callaway et al., 2022), and decision making (Lieder & Griffiths, 2017; Hahn & Warren, 2009). The ideas emphasized in the thesis, such as mental simulation (Chapter 5), resource rational decision making (Chapter 6), and generalization (Chapter 7), are not limited to the current setting. They resonate throughout various inquiries under the domain of model-based cognition. Temporal causal learning is promising research topic, as it has the potential to shed light on and improve our understanding of the complexities of model-based cognition in the human mind.

Evidence processing For decades, cognitive researchers wonder how people learn so much from so little, i.e. how one acquire substantial knowledge from limited information. Multiple perspectives have contributed valuable insights to this inquiry. These perspectives underscore the presence of cognitive mechanisms enabling the integration of prior beliefs (Griffiths & Tenenbaum, 2009) and engagement in hierarchical and compositional structures (Ullman et al., 2018; Lake et al., 2017). Nevertheless, this thesis presents one additional explanation: Perhaps the information encapsulated within the data is not inherently sparse, but rather rich when considering the temporal dimension, especially under the fact that we have cognitive means to effectively utilize it. By emphasizing the significance of temporal information, this thesis contributes to the understanding of how individuals extract and leverage valuable information from seemingly limited data. In reality, people frequently learn from direct perception. It will become more and more important for us to study and understand how information is extracted from these perceptual experiences.

Natural cognition One motivation behind this thesis is to gain insights into natural cognition as it unfolds in everyday contexts (Loftus, 1981). Incorporating the element of time brings the

causal learning task a step closer to mirroring real-world scenarios. The empirical studies in Chapter 5 and Chapter 6 underscore the resonance between many of these findings and those observed in previous atemporal studies, especially regarding how people are sensitive to the base rate, how generative and preventative causes can intertwine in influencing people's judgments, and what kind of local errors people tended to make when learn causal structures. The empirical studies in Chapter 5 and Chapter 6 underscore the resonance between many of these findings and those observed in previous atemporal studies, particularly concerning people's sensitivity to the base rate (Rottman, 2016), the interplay of generative and preventative causes in influencing people's judgments (Buehner et al., 2003; Cheng, 1997), and the types of local errors individuals tended to make when learning causal structures (Fernbach & Sloman, 2009; Davis et al., 2020). However, they also can offer some insights into puzzles that cannot be fully resolved through atemporal causal learning studies. For instance, in Chapter 6, I demonstrate that with temporal information considered, intervening on the root node could yield more informative results than intervening on the non-root node, as long as the agent can process the abundant information provided.

This is very different from the atemporal setting where intervening on the root node can often introduce confounding due to the spontaneous activation of all direct or indirect effects stemming from the root node (Steyvers et al., 2003; Coenen et al., 2015). Researchers find it difficult to explain the root-node preference (i.e. postive testing Coenen et al., 2015) observed when people tend to intervene on the root node in the atemporal setting, while the temporal context introduced in this thesis presents one potential explanation: People who possess familiarity with temporal causal learning might hold the belief that root node interventions encompass more comprehensive information. Such a belief may carry over to the atemporal context, which is less familiar in their daily lives. [1] Thus, by studying causal reasoning within more natural settings, we can anticipate uncovering explanations for specific behavioral tendencies exhibited by individuals.

Interdisciplinary contributions Studying the mental process of causal learning is important, as causal discovery stands not only as an essential cognitive process but also as the very foundation of science itself. The scientific community has exerted collective effort to develop methodologies for causal discovery. Meanwhile, this topic is of great significance for computer scientists who strive to create intelligent tools or facilitate scientific breakthroughs. Thus, causal cognition occupies a central position within the triad of individual, artificial, and collective intelligence. This thesis is anchored in the domain of individual intelligence, yet it casts illumination on other dimensions as well. For instance, in Chapter 5 and Chapter 6, I describe the aspects that individuals harness to enhance the efficiency and computational economy of causal learning. These insights could prove invaluable in shaping the development of human-like artificial intelligence. Furthermore, Chapter 7 demonstrates that even when presented with the same empirical

[1] I am grateful to Dave Lagnado for raising this point.

data, laypeople's causal judgments can be influenced by instructions detailing the decision-making processes scientists employed during data collection. This underscores the importance of context in scientific communication.

8.3 A roadmap of future topics

Time and causality, is a field that is still in its infancy, awaiting further exploration. I have discussed various future directions at the end of Chapter 5-7, each tailored to harmonize with the empirical inquiries in their respective chapters. However, moving beyond the specific empirical inquiries tackled within this thesis, there are broader horizons to consider within the domain of time and causality. I will here discuss topics that may not be intrinsically linked to the empirical questions presented herein. Instead, they serve to underscore a broader and more comprehensive interest.

8.3.1 More forms of causation

So far, causal learning studies, including this thesis, have focused on learning about generative and preventative relationships. It is noted that there are many other forms of causal relationships, especially after considering time.

Hasten and delay Hasten and delay are two kinds of causal relationships that merely alter the timing of effects without affecting their frequency (Bennett, 1987). In other words, they do not generate or prevent events but instead modify *when* they occur.

When we have the means to track individual effects, it is not hard to distinguish between generation/prevention and hastening/delaying. For instance, if we assign a unique identifier to each bus sharing the same number, we can determine whether a bus is generated (new) or hastened when it arrives unexpectedly early in a morning. In contrast, when we have no access to the individual identifications, hasten and delay share similarities with generation and prevention: They both lead to an increase or decrease in the number of outcomes over the short term. The bacteria task from Greville & Buehner (2007), introduced in Chapter 4 and 7, can arguably be viewed through the lens of hasten and delay. This is because the outcome, "death", is inevitable for each sample. However, given that the records only provided a macro-level information about daily death numbers, we can still employ the rate-based model, without explicitly categorizing it as generative/preventative or hastening/delaying.

People may also prefer to use the word "generation" and "hasten" in different situations. Taking the bacteria task as an example again, if the treatment has an age-dependent effect that causes each sample to die two days earlier than its originally expected time of death, we might consider

it a hastening effect. Conversely, if the treatment indifferently causes a large sample to die simultaneously, we might say it generates death or "kills" the samples.

Regularization Regularization is another potentially intriguing form of causation. In this case, the cause does not necessarily generate or prevent the effect but rather *maintains it at a specific level*. Under point event circumstances, we can describe this maintenance as occurring at a particular rate. This particular nonlinear situation may involve complex causal mechanisms. Similar to preventative causation, which necessitates base rate activations, regularization requires the system to undergo dynamic changes for us to discern the influence of the regularization cause or causes. Confirming the role of regularization can be challenging since the cause often appears to have no discernible effect. Many biological systems, including the human body, operate as complex regularization systems. Consequently, understanding the function of regularization becomes crucial. It frequently demands careful observation of system dynamics (Ross, 2015).

8.3.2 Causality and temporal perceptions

This thesis explores the impact of perceived temporal information on the assessment of relationships between variables. Additionally, in Chapter 3, I review studies that highlight how knowledge of causal relationships can influence the perception of temporal delay and temporal order (Hoerl et al., 2020; Bechlivanidis & Lagnado, 2013; Bechlivanidis et al., 2022).

Prior The perception of causality and the perception of time can influence each other. This bidirectional influence stems from the inherent uncertainty associated with both domains. Causal induction, being an inductive problem, is inherently uncertain. Recent research has shown that individuals can provide systematic and predictable ratings about the confidence of the causal judgment alongside the causal judgment itself (O'Neill, Henne, Bello, et al., 2022; O'Neill, Henne, Pearson, & De Brigard, 2022). Concurrently, human perceptions are inherently noisy, and the input can vary in ambiguity, which could further depend on the modality (e.g. auditory vs. visual B. C. Moore, 2012; Kanabus et al., 2002). Following a Bayesian framework, when evidence is uncertain, individuals tend to rely on their prior knowledge to make judgments (De Lange et al., 2018; Seriès & Seitz, 2013). In this thesis, a uniform prior assumption for different causal structures is often employed. However, in more realistic settings, individuals are likely to apply their prior knowledge when learning causal relationships, particularly in continuous time scenarios where evidence could be more extensive and harder to process than in atemporal settings (see Btesh et al., 2023).

Temporal perceptions as causal events How people perceive and measure time remains an unsolved question (Buonomano, 2017). One hypothesis suggests that individuals gauge time based on the number of events they experience. For instance, your may feel that time appears

to fly by when exploring a new place during a vacation. However, a month later, you may have a contrasting feeling that a single day of tourism was longer compared to a typical working day (Buonomano, 2017). This discrepancy can be attributed to differences between predictive and retrospective temporal estimation. When estimating time retrospectively, people often consider the number of events experienced (Jones, 2019). Thus, events serve as important cues for temporal duration estimation.

Meanwhile, imagining the causal chain of events could help estimate the temporal duration between two events. For example, when waiting for an online order, one might imagine the process of the order being sent to the store, packaged, and dispatched (Buchanan et al., 2010). Similarly, when anticipating the effects of medication, one might imagine the pill dissolving in the stomach, entering the bloodstream, and reaching the intended target area. Stephan, Tentori, et al. (2021) discovered that having more knowledge about the detailed mechanism reduces the perception of a causal relationship between two variables. One possible explanation is that individuals confuse a preexisting yet unknown mechanism with a newly discovered mechanism that could serve as a probabilistic mediation. Building upon this thesis, we can propose another possibility: By considering the details of intermediate mechanisms, reasoners may inadvertently extend the perceived delays between cause and effect, subsequently reducing the perceived causal relationship between the two variables.

8.3.3 Continuous values vs. point events

This thesis primarily focuses on how individuals infer causal relationships from point events in continuous time, whereas other studies have examined how people infer from continuous values in continuous time (Davis et al., 2020; Rehder et al., 2022; Soo & Rottman, 2018; Zhang & Rottman, 2021b; Btesh et al., 2023). However, rather than viewing these approaches as separate mechanisms, they are likely to form a hierarchical structure to explain causal phenomena. To illustrate this, consider the predator-prey relationship between fish and weed. At the lower level, we can examine individual events such as a fish consuming a plant. Moving to a higher level, we can analyze how the *populations* of fish and weed change over time, based on continuous values. Finally, at an even higher level, we can investigate how each species experiences cyclic patterns of *bloom* and *die out*, which pertains to the event level again. By moving between these hierarchical levels, we can switch between continuous values and events.

Importantly, it is not necessary to limit our thinking to the level provided by the data. For example, in Chapter 4, the rate-based model outperformed the event-based model in capturing human performance in identifying cyclic structures. This suggests that when the number of events is large, individuals may employ a relatively continuous representation (i.e. the rate). Similarly, in Chapter 5, I demonstrate that summary statistics, such as counting the occurrences of effects following the interventions, can be valuable when direct inference from raw events is challenging.

Additionally, Rehder et al. (2022) showed that in a setting involving continuous values, individuals may abstract "events" from trends by focusing on obvious moments of increase or decrease while overlooking micro-dynamics. These findings indicate that people have the ability to use cognitively plausible representations and abstract higher-level features from the data.

8.3.4 Laypeople's theories of causal learning

Research on causal induction has predominantly focused on controlled scientific settings, but in this thesis, I emphasize the importance of considering the data people encounter in their everyday lives, which may differ from laboratory conditions. Real-world data often involves temporal information and interventions that occur over time, introducing interdependencies between data points. To gain a comprehensive understanding of laypeople's theories of causal learning, it is crucial to study causal learning in more ecological situations that reflect the complexity and dynamics of real-life scenarios.

Beyond randomized experiments In the early stages of causal discovery, before the development of formal tools and technological advancements, humans did not have access to scientific methods such as randomized experiments (Athey & Imbens, 2017), causal graphical models (Pearl, 2000), instrumental variables (Imbens, 2014), etc. How did people navigate the realm of causal discovery without these tools? For instance, people from different cultures independently invented calendars that align with the rules governing the Earth's orbit around the sun and the moon's orbit around the Earth (Duncan, 1999). People can use herbal remedies to treat illnesses, even in the absence of randomized medical trials. [2] Their methods may often involved small samples, repeated measurements, and temporal information (Carlisle & Eldar, 2005; Sternberg et al., 2001). This thesis focuses on individual causal discovery within a short time scale. However, it is equally important, if not more so, to explore causal beliefs on a larger scale, i.e. conventions related to causal beliefs that are not solely confirmed by well-acknowledged scientific methods (Dubova & Goldstone, 2023). By comparing the so-called "scientific approach" with the approaches adopted by laypeople, we can gain insights into how different cultures and conventions have emerged over generations in the domain of causal discovery.

Time travel and causality Time travel has become an increasingly popular topic in science fiction (Dubourg & Baumard, 2022). These stories often feature characters who can traverse time, make changes, and then return to the present. However, readers, myself included, sometimes find these narratives challenging to grasp, especially when authors introduce changes at the beginning of the story, before the time travel occurs, creating what is known as a closed loop (Suddendorf & Corballis, 1997; Smith, 2018).

[2] Thanks to my roommate Elva Peng for inspiring me to think about it.

Two classic philosophical views on time exist: presentism and eternalism. Presentism regards the present as the sole reality of time, with the past having already transpired and the future yet to unfold. In contrast, eternalism does not differentiate between the past, present, and future (Buonomano, 2017). The difficulty in comprehending time travel may reflect a presentist perspective. Moreover, the challenge often lies in understanding the causal order of events within the closed loop, as cause and effect can become intertwined. This suggests a profound connection between causality and our understanding of time, which could be investigated in the future.

Causal predictions and actions Learning causal structures with temporal information is not the final step in the cognitive process. We must also utilize the causal structures we have learned to predict future outcomes and take timely actions. This applies not only to daily life but also to various forms of art. For instance, in music, skilled composers strive to create dynamic melodies that are both unfamiliar and yet conform to conventional patterns, striking a balance between expectation and surprise (Levitin, 2006). In stand-up comedy, the timing of delivering a punchline is crucial. Comedians wait for the audience to anticipate what comes next and then deliver a punchline that subverts their expectations to some degree (Buonomano, 2017; Martin & Ford, 2018). In photography and filmmaking, the arrangement and sequencing of photos or scenes can profoundly alter the impact of the narrative, even when using the same materials. Future research could explore how people employ causal models to guide their actions, particularly in situations where performers must consider the causal models of their audience or receptors.

8.4 Conclusion

This thesis explores the process of human causal structure induction from events in continuous time. It builds a bridge between the computational and algorithmic levels in causal reasoning, and provide quantitative predictions about human judgments in various situations. By understanding the mechanisms behind people's rapid and efficient learning with limited resources, this thesis contributes to our understanding of natural cognition while also offering insights into the quest for more human-like algorithms. In our daily lives, we encounter not only expected or surprising events but also ponder their connections to the past and future. Therefore, it is crucial to incorporate a formal framework for temporal causal inference into the theory of human causal reasoning.

References

Ahn, W.-k., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*(3), 299–352.

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.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, *114*(3), 435–448.

Allan, L. G., & Jenkins, H. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14*(4), 381–405.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Psychology Press. doi: 10.4324/9780203771730

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., ... Orben, A. (2022). Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*.

Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments* (Vol. 1, pp. 73–140). Elsevier.

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*(3), 499–526. doi: 10.1111/j.1551-6709.2010.01161.x

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

Baker, A. G., Berbrier, M. W., & Vallee-Tourangeau, F. (1989). Judgements of a 2× 2 contingency table: Sequential processing and the learning curve. *The Quarterly Journal of Experimental Psychology Section B*, *41*(1b), 65–97.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413–424.

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.

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.

Bechlivanidis, C., & Lagnado, D. A. (2013). Does the "why" tell us the "when"? *Psychological Science*, *24*(8), 1563–1572.

Bechlivanidis, C., & Lagnado, D. A. (2016). Time reordered: Causal perception guides the interpretation of temporal order. *Cognition*, *146*, 58–66.

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 238–249.

Bennett, J. (1987). Event causation: The counterfactual analysis. *Philosophical Perspectives*, *1*, 367–386.

Biele, G., Erev, I., & Ert, E. (2009). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of mathematical psychology*, *53*(3), 155–167.

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2022). Heuristics from bounded meta-learned inference. *Psychological Review*, *129*(5), 1042–1077.

Blakey, E., Tecwyn, E. C., McCormack, T., Lagnado, D. A., Hoerl, C., Lorimer, S., & Buehner, M. J. (2019). When causality shapes the experience of time: Evidence for temporal binding in young children. *Developmental Science*, *22*(3), e12769.

Blum, M. G., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science*, *28*(2), 189–208.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive Psychology*, *74*, 35–65. doi: 10 .1016/j.cogpsych.2014.06.003

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338. doi: 10.1037/ rev0000061

Bramley, N. R., Gerstenberg, T., & Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 236–241).

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. doi: 10.1037/xlm0000548

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2019). Intervening in time. In S. Kleinberg (Ed.), *Time and causality across the sciences* (pp. 86–115). Cambridge University Press.

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *195*, 9–38. doi: 10.1016/j.cogpsych.2018.05.001

Bramley, N. R., Jones, A., Gureckis, T. M., & Ruggeri, A. (2022). Childrenâ€™s failure to control variables may reflect adaptive decision-making. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-022-02120 -1

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731. doi: 10.1037/xlm0000061

Bramley, N. R., Mayrhofer, R., Gerstenberg, T., & Lagnado, D. A. (2017). Causal learning from interventions and dynamics in continuous time. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 150–155).

Brändle, F., Stocks, L. J., Tenenbaum, J. B., Gershman, S. J., & Schulz, E. (2023). Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. doi: 10.1037/met0000159

Btesh, V. J., Bramley, N., Speekenbrink, M., & Lagnado, D. (2023). Less is more: Adaptive strategies in continuous time causal learning. *PsyArXiv*.

Buchanan, D., Tenenbaum, J., & Sobel, D. (2010). Edge replacement and nonindependence in causation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 919–924).

Buehner, M. J. (2006). A causal power approach to learning with rates. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 28).

Buehner, M. J. (2012). Understanding the past, predicting the future: causation, not intentional action, is the root of temporal binding. *Psychological Science*, *23*(12), 1490–1497.

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), 1119–1140. doi: 10.1037/0278-7393.29.6.1119

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning*, *8*(4), 269–295.

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., & May, J. (2004). Abolishing the effect of reinforcement delay on human causal learning. *Quarterly Journal of Experimental Psychology Section B*, *57*(2), 179–191. doi: 10.1080/02724990344000123

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. doi: 10.1080/13546780500368965

Buonomano, D. (2017). *Your brain is a time machine: The neuroscience and physics of time*. New York: W.W. Norton & Company.

Burns, P., & McCormack, T. (2009). Temporal information and children's and adults' causal inferences. *Thinking & Reasoning*, *15*(2), 167–196. doi: 10.1080/13546780902743609

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*. doi: 10.1038/s41562-022-01332-8

Carlisle, E., & Eldar, S. (2005). Heuristics and biases in attitudes towards herbal medicines. In P. N. J.-L. V. Girotto (Ed.), *The shape of reason* (pp. 221–240). London: Psychology Press.

Carroll, C. D., & Cheng, P. (2009). Preventative scope in causation. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 833–838).

Carroll, C. D., & Kemp, C. (2015). Evaluating the inverse reasoning account of object discovery. *Cognition*, *139*, 130–153.

Carroll, S. M. (2008). The cosmic origins of time's arrow. *Scientific American*, *298*(6), 48–57.

Carroll, S. M. (2022). *The biggest ideas in the universe: Space, time, and motion*. New York: Dutton.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120. doi: 10.1111/1467-8624.00081

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

Cheng, P. W., & Lu, H. (2017). Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing. In M. Waldmann (Ed.), *The oxford handbook of causal reasoning* (p. 65-84). New York: Oxford University Press.

Chow, J. Y., Lee, J. C., & Lovibond, P. F. (2023). Inhibitory learning with bidirectional outcomes: Prevention learning or causal learning in the opposite direction? *Journal of Cognition*, *6*(1), 1-24.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, 1–72. doi: 10.1017/S0140525X1500031X

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. doi: 10.3758/s13423-018-1470-5

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133. doi: 10.1016/j.cogpsych.2015.02.004

Coenen, A., Ruggeri, A., Bramley, N. R., & Gureckis, T. M. (2019). Testing one or multiple: How beliefs about sparsity affect causal experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(11), 1923–1941. doi: 10.1037/xlm0000680

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate bayesian computation (abc) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25. doi: 10.1016/j.cogpsych.2017.05.001

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412–441.

Davis, Z., Bramley, N. R., & Rehder, B. (2020). Causal structure learning in continuous systems. *Frontiers in Psychology*, *11*, 244. doi: 10.3389/fpsyg.2020.00244

Davis, Z., Bramley, N. R., Rehder, B., & Gureckis, T. M. (2018). A causal model approach to dynamic control. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 281–286).

Davis, Z., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*(5), e12839.

Dean, T., & Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, *5*(2), 142–150.

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779.

Derringer, C., & Rottman, B. M. (2018). How people learn about causal influence when there are many possible causes: A model based on informative transitions. *Cognitive Psychology*, *102*, 41–71. doi: 10.1016/j.cogpsych.2018.01.002

Dubourg, E., & Baumard, N. (2022). Why imaginary worlds? the psychological foundations and cultural evolution of fictions with imaginary worlds. *Behavioral and Brain Sciences*, *45*, e276.

Dubova, M., & Goldstone, R. L. (2023). Carving joints into nature: reengineering scientific concepts in light of concept-laden evidence. *Trends in Cognitive Sciences*, *27*(7), 656–670.

Duncan, D. E. (1999). *Calendar:: Humanity's epic struggle to determine a true and accurate year*. Harper Collins.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmutz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: John Wiley & Sons.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3–19.

Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., ... Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 678–693. doi: 10.1037/a0014928

Fraser, K. M., & Holland, P. C. (2019). Occasion setting. *Behavioral Neuroscience*, *133*(2), 145-175.

Garcia, J., Ervin, F. R., & Kölling, R. A. (1966). Learning with prolonged delay of reinforcement. *Psychonomic Science*, *5*(3), 121–122.

Gershman, S. J. (2020). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, 104394. doi: 10.1016/j.cognition.2020.104394

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

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), 936â€“975. doi: 10.1037/rev0000281

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744.

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, 104842.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Gong, T., & Bramley, N. R. (2020). What you didn't see: Prevention and generation in continuous time causal induction. In S. Denison, M. Mack, Y. Xu, & B. Armstrong (Eds.), *Proceedings of the 42th annual conference of the cognitive science society* (pp. 2908–2914).

Gong, T., & Bramley, N. R. (2022). Intuitions and perceptual constraints on causal learning from dynamics. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th annual conference of the cognitive science society* (pp. 1455–1461).

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

Gong, T., & Bramley, N. R. (2023b). Evidence from the future. *PsyArXiv (Accepted at Journal of Experimental Psychology: General)*.

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

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–119. doi: 10.1037/a0021336

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*(2), 87–92.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*(5), 620–629. doi: 10.1037/0012-1649.37.5.620

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, bayesian learning and cognitive development. *Developmental science*, *10*(3), 281–287.

Goudie, R. J., & Mukherjee, S. (2011). *An efficient gibbs sampler for structural inference in bayesian networks* (Tech. Rep.). Citeseer.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.

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. doi: 10.1037/a0020976

Greville, W. J., & Buehner, M. J. (2016). Temporal predictability enhances judgements of causality in elemental causal induction from both observation and intervention. *Quarterly Journal of Experimental Psychology*, *69*(4), 678–697.

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. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, *24*(11), 873–883.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. doi: 10.1111/tops.12142

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

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661-716. doi: 10.1037/a0017201

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.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Guillory, A. (2012). *Active learning and submodular functions.* University of Washington.

Hagmayer, Y., Meder, B., Osman, M., Mangold, S., & Lagnado, D. (2010). Spontaneous causal learning while controlling a dynamic system. *The Open Psychology Journal*, *3*(1), 145–162.

Hagmayer, Y., & Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition*, *30*(7), 1128–1137. doi: 10.3758/BF03194330

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

Hall, G., & Honey, R. C. (1989). Contextual effects in conditioning, latent inhibition, and habituation: Associative and retrieval functions of contextual cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *15*(3), 232.

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

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Harman, G. (1986). *Change in view: Principles of reasoning*. The MIT Press.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science*, *31*(5), 765–814.

Hawkins, R. D., Gweon, H., & Goodman, N. D. (2021). The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive Science*, *45*(3), e12926. doi: 10.1111/cogs.12926

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

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

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, *268*(5214), 1158–1161.

Hoerl, C., Lorimer, S., McCormack, T., Lagnado, D. A., Blakey, E., Tecwyn, E. C., & Buehner, M. J. (2020). Temporal binding, causation, and agency: Developing a new theoretical framework. *Cognitive Science*, *44*(5), e12843.

Hume, D. (1740). *A treatise of human nature*. New York: Oxford University Press (2000 reprint).

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Imbens, G. (2014). *Instrumental variables: an econometrician's perspective* (Tech. Rep.). National Bureau of Economic Research.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological monographs: General and Applied*, *79*(1), 1.

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

Jones, L. A. (2019). The perception of duration and the judgment of the passage of time. In V. Arstila, A. Bardon, S. E. Power, & A. Vatakis (Eds.), *The illusions of time: Philosophical and psychological essays on timing and time perception* (p. 53-67). London: Palgrave Macmillan.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus & Giroux.

Kamin, L. J. (1967). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton-Century-Crofts.

Kanabus, M., Szelag, E., Rojek, E., & Poppel, E. (2002). Temporal order judgement for auditory and visual stimuli. *Acta Neurobiologiae Experimentalis*, *62*(4), 263–270.

Kim, N. S., & Ahn, W.-k. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General*, *131*(4), 451–476. doi: 10.1037/0096-3445.131.4.451

Klayman, J., & Ha, Y.-w. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 596–604. doi: 10.1037/0278-7393.15.4.596

Kuhn, D., & Brannock, J. (1977). Development of the isolation of variables scheme in experimental and" natural experiment" contexts. *Developmental Psychology*, *13*(1), 9–14. doi: 10.1037/0012-1649.13.1.9

Lagnado, D. A. (2011). Causal thinking. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 129–149). New York: Oxford University Press.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 1036–1073.

Lagnado, D. A., & Sloman, S. (2002). Learning causal structure. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th annual meeting of the cognitive science society* (pp. 560–565).

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856–876. doi: 10.1037/0278-7393.30.4.856

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. doi: 10.1037/0278-7393.32.3.451

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

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). New York: Oxford University Press.

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. *The Psychology of Learning and Motivation*, 195–227. doi: 10.1016/bs.plm.2021.02.004

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Lee, J. C., & Lovibond, P. F. (2021). Individual differences in causal structures inferred during feature negative learning. *Quarterly Journal of Experimental Psychology*, *74*(1), 150–165.

Le Pelley, M. E., Griffiths, O., & Beesley, T. (2017). Associative accounts of causal cognition. In M. Waldmann (Ed.), *The oxford handbook of causal reasoning* (p. 13-28). New York: Oxford University Press.

Le Pelley, M. E., & McLaren, I. P. L. (2001). Retrospective revaluation in humans: Learning or memory? *The Quarterly Journal of Experimental Psychology Section B*, *54*(4b), 311–352.

Levitin, D. J. (2006). *This is your brain on music: The science of a human obsession.* London: Penguin.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762–794.

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. doi: 10.1017/S0140525X1900061X

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1–32.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322–349. doi: 10.3758/s13423-017-1286-8

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2017). Fundamentals and recent developments in approximate bayesian computation. *Systematic Biology*, *66*(1), e66–e82.

Lipp, O. V., & Vaitl, D. (1992). Latent inhibition in human pavlovian differential conditioning: Effect of additional stimulation after preexposure and relation to schizotypal traits. *Personality and Individual Differences*, *13*(9), 1003–1012.

Loftus, E. F. (1981). Natural and unnatural cognition. *Cognition*, *10*(1-3), 193–196.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Lovibond, P. F., & Lee, J. C. (2021). Inhibitory causal structures in serial and simultaneous feature negative learning. *Quarterly Journal of Experimental Psychology*, *74*(12), 2165–2181.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955. doi: 10.1037/a0013256

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299. doi: 10.1016/j.cognition.2013.12.010

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147. doi: 10.1111/j.1551-6709.2009.01058.x

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.

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

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602–1611.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, *127*, 101396.

Malthus, T. R. (1872). *An essay on the principle of population*. Reeves & Turner.

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive Science*, *40*(1), 100–120.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge: MIT Press. doi: 10.7551/mitpress/9780262514620.001.0001

Martin, R. A., & Ford, T. (2018). *The psychology of humor: An integrative approach*. Cambridge: Academic Press.

McCormack, T., Bramley, N. R., Frosch, C., Patrick, F., & Lagnado, D. (2016). Childrenâ€™s use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22. doi: 10.1016/j.jecp.2015.06.017

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

Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and intervening: Rational and heuristic models of causal decision making. *The Open Psychology Journal*, *3*, 119–135. doi: 10

Melchers, K. G., Wolff, S., & Lachnit, H. (2006). Extinction of conditioned inhibition through nonreinforced presentation of the inhibitor. *Psychonomic Bulletin & Review*, *13*(4), 662–667.

Mendelson, R., & Shultz, T. R. (1976). Covariation and temporal contiguity as principles of causal inference in young children. *Journal of Experimental Child Psychology*, *22*(3), 408–412.

Meng, Y., Bramley, N. R., & Xu, F. (2018). Childrenâ€™s causal interventions combine discrimination and confirmation. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society.*

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Moore, B. C. (2012). *An introduction to the psychology of hearing.* Brill.

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: a review. *Consciousness and Cognition*, *21*(1), 546–561.

Nelson, J. D. (2005). Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–1000. doi: 10.1037/0033-295X.112.4.979

Nikolic, M., & Lagnado, D. A. (2015). There aren't plenty more fish in the sea: A causal network approach. *British Journal of Psychology*, *106*(4), 564–582. doi: 10.1111/bjop.12113

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* New York: Oxford University Press.

O'Neill, K., Henne, P., Bello, P., Pearson, J., & De Brigard, F. (2022). Confidence and gradation in causal judgment. *Cognition*, *223*, 105036.

O'Neill, K., Henne, P., Pearson, J., & De Brigard, F. (2022). Measuring and modeling confidence in human causal judgment. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th annual conference of the cognitive science society* (pp. 446–452).

Pacer, M. (2016). *Mind as theory engine: causation, explanation and time* (Unpublished doctoral dissertation). UC Berkeley.

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).

Pacer, M., & Griffiths, T. L. (2015). Upsetting the contingency table: Causal induction over sequences of point events. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 1805–1810).

Paul, L. A., & Hall, N. (2013). *Causation: A user's guide.* Oxford University Press.

Pavlov, I. P. (1928). *Lectures on conditioned reflexes.* New York: W H Gantt International Publishers.

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, *52*(1), 111–139.

Pearl, J. (2000). *Causality.* New York: Cambridge University Press (2009 reprint). doi: 10.1017/CBO9780511803161

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect.* New York: Basic Books.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review*, *14*, 577–596.

Petitet, P., Attaallah, B., Manohar, S. G., & Husain, M. (2021). The computational cost of active information sampling before decision-making under uncertainty. *Nature Human Behaviour*, *5*, 935–946. doi: 10.1038/s41562-021-01116-6

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346–354.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness, and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Quine, W. V. O. (1960). *Word and object*. Cambridge: MIT press.

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*(3), 784–805.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107. doi: 10.1016/j.cogpsych.2014.02.002

Rehder, B. (2017). Reasoning with causal cycles. *Cognitive Science*, *41*, 944–1002. doi: 10.1111/cogs.12447

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

Rescorla, R. A., & Wagner, A. R. (1972). A theory on pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning ii: Current theory and research* (pp. 64–99). New York: Appleton Century Crofts.

Ross, L. N. (2015). Dynamical models and explanation in neuroscience. *Philosophy of Science*, *82*(1), 32–54.

Rottman, B. M. (2016). Searching for the best cause: Roles of mechanism beliefs, autocorrelation, and exploitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1233–1256.

Rottman, B. M. (2017). The acquisition and use of causal structure knowledge. In M. Waldmann (Ed.), *The oxford handbook of causal reasoning* (pp. 85–114). New York: Oxford University Press.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*(1), 109-139. doi: 10.1037/a0031903

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology*, *87*, 88–134. doi: 10.1016/j.cogpsych.2016.05.002

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*(1-2), 93–125. doi: 10.1016/j.cogpsych.2011.10.003

Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning and Motivation*, *6*(3), 314–326.

Schlottmann, A., Cole, K., Watts, R., & White, M. (2013). Domain-specific perceptual causality in children depends on the spatio-temporal configuration, not motion onset. *Frontiers in psychology*, *4*, 365.

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

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322–332. doi: 10.1111/j.1467-7687.2007.00587.x

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, *7*, 668.

Settles, B. (2009). *Active learning literature survey.* Technical Report University of Wisconsin-Madison.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1–21.

Shanks, D. R. (1995). *The psychology of associative learning.* Cambridge University Press.

Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition*, *19*(4), 353–360.

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.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*, 99–124. doi: 10.1146/annurev-neuro-072116-031526

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

Simon, H. A. (1982). *Models of bounded rationality: Empirically grounded economic reason.* MIT press. doi: 10.7551/mitpress/4711.001.0001

Skinner, B. F. (1938). *The behaviour of organisms: An experimental analysis.* New York: D. Appleton-Century Company Incorporated.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

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

Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228. doi: 10.1016/S0364-0213(99)80039-1

Smith, N. J. (2018). *Time travel. stanford encyclopedia of philosophy.* Retrieved from `https://plato.stanford.edu/entries/time-travel`

Sobel, D. M., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory & Cognition*, *34*(2), 411–419. doi: 10.3758/BF03193418

Soo, K. W., & Rottman, B. M. (2018). Causal strength induction from time series data. *Journal of Experimental Psychology: General*, *147*(4), 485–513. doi: 10.1037/xge0000423

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.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search* (2nd ed.). MIT press. doi: 10.1007/978-1-4612-2748-9

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation–a computational model. *Cognitive Science*, *44*(7), e12871. doi: 10.1111/cogs.12871

Stephan, S., Placì, S., & Waldmann, M. R. (2021). Evaluating general versus singular causal prevention. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43th annual conference of the cognitive science society* (pp. 1402–1408).

Stephan, S., Tentori, K., Pighin, S., & Waldmann, M. R. (2021). Interpolating causal mechanisms: The paradox of knowing more. *Journal of Experimental Psychology: General*, *150*(8), 1500–1527. doi: 10.1037/xge0001016

Sternberg, R. J., Nokes, C., Geissler, P. W., Prince, R., Okatcha, F., Bundy, D. A., & Grigorenko, E. L. (2001). The relationship between academic and practical intelligence: A case study in kenya. *Intelligence*, *29*(5), 401–418.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489. doi: 10.1207/s15516709cog2703_6

Suddendorf, T., & Corballis, M. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs*, *123*(2), 133–167.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate bayesian computation. *PLoS Computational Biology*, *9*(1), e1002803.

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.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.

Tarpy, R. M., & Sawabini, F. L. (1974). Reinforcement delay: A selective review of the last decade. *Psychological Bulletin*, *81*(12), 984–997.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410. doi: 10.1037/rev0000052

Tong, S., & Koller, D. (2001). Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence* (Vol. 17, pp. 863–869).

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82. doi: 10.1016/j.cogpsych.2017.05.006

Valentin, S., Bramley, N. R., & Lucas, C. G. (2020). Learning hidden causal structure from temporal data. In S. Denison, M. Mack, Y. Xu, & B. Armstrong (Eds.), *Proceedings of the 42th annual conference of the cognitive science society* (pp. 1906–1912).

Valentin, S., Bramley, N. R., & Lucas, C. G. (2022). Discovering common hidden causes in sequences of events. *Computational Brain & Behavior*, 1–23.

Van Rooij, I., Blokpoel, M., Kwisthout, J., & Wareham, T. (2019). *Cognition and intractability: A guide to classical and parameterized complexity analysis*. Cambridge University Press. doi: 10.1017/9781107358331

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637. doi: 10.1111/cogs.12101

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 53–76.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216–227.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222–236.

Wasserman, E. A., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgement: The role of within-compound associations. *The Quarterly Journal of Experimental Psychology: Section B*, *51*(2), 121–138.

White, P. A. (2006). How well is causal structure inferred from cooccurrence information? *European Journal of Cognitive Psychology*, *18*(3), 454–480. doi: 10.1080/09541440500264861

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, *25*(A), 287–298.

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

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science*, *10*(2), 92–97.

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.

Zhang, Y., & Rottman, B. (2021b). Causal learning with interrupted time series. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43th annual conference of the cognitive science society* (pp. 1333–1339).

Zhang, Y., & Rottman, B. M. (2021a). Causal learning with delays up to 21 hours. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43th annual conference of the cognitive science society* (pp. 2766–2772).

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.

Zhao, Y., Zeng, T., Wang, T., Fang, F., Pan, Y., & Jia, J. (2023). Subcortical encoding of summary statistics in humans. *Cognition*, *234*, 105384.

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.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. doi: 10.1016/j.dr.2006.12.001

Appendices

Appendix A

Appendices for Chapter 5

A.1 Normative calculations

The normative learner updates the prior over structures $P(S)$ (here assumed to be uniform), with a likelihood function to obtain a posterior distribution, given the set of gamma parameters $\mathbf{w}$ which indicates the belief about delays:

$$P(S|\mathbf{d}, \mathbf{w}; \mathbf{i}) \propto p(\mathbf{d}|S, \mathbf{w}; \mathbf{i}) \cdot P(S) \tag{A.1}$$

Here $\mathbf{d}$ refers to effect data (E's activations), which is conditioned upon a set of interventions $\mathbf{i}$ on the causes (A or B).

In order to maintain rational beliefs about causal structure, the ideal reasoner considers all possible causal paths $\mathbf{Z}_s$ that could describe what actually happened given each possible structural hypothesis $s \in \mathbf{S}$, summing up the individual likelihood of these mutually exclusive and exhaustive possibilities to assess the overall likelihood of each structure hypothesis:

$$P(\mathbf{d}|s, \mathbf{w}; \mathbf{i}) = \sum_{z' \in \mathbf{Z}_s} P(z'|s, \mathbf{w}; \mathbf{i}) \tag{A.2}$$

Normative causal attribution involves three steps: 1) attributing causes to effects that have occurred; 2) explaining away effects that should or might have occurred but were not observed; 3) examining the temporal distance between presumed preventative events and the subsequent effect event. Step 1 and 2 correspond to path construction. We use $\{\alpha_g, \beta_g\}$,$\{\alpha_p, \beta_p\}$, $\{\alpha_b, \beta_b\}$ to denote parameters of gamma distributions for generative delays, preventative windows, and base rate delays. In the current experiments: $\{\alpha_g = 9, \beta_g = 6\}$, $\{\alpha_p = 36, \beta_p = 12\}$, and $\{\alpha_b = 100, \beta_b = 20\}$ (regular base rate) or $\{\alpha_b = 1, \beta_b = 0.2\}$ (irregular base rate).

Step 1 is to form $g' \rightarrow e'$ pairs where 1) the effect event e' is not over-determined (i.e. has a single actual cause), 2) the cause event g' does not produce its effect twice, and 3) g' precedes

e'. The likelihood of each pair is then determined by mapping the delay between g' and e' to the gamma density function:

$$P(g' \to e'|\alpha_g, \beta_g) = P(t_{g' \to e'} = t_{g'e'}|\alpha_g, \beta_g) \tag{A.3}$$

Step 2 involves forming $g' \to h$ pairs where h is a hidden effect event assumed to happen sometime after the observable period *or* at some point during a preventative window. The likelihood calculation depends on the gamma cumulative density falling beyond the end of the clip or within the window:

$$\begin{aligned} P(g' \to h|\alpha_g, \beta_g, \alpha_p, \beta_p) = P(t_{g' \to h} > t_{end}|\alpha_g, \beta_g) + \\ P(t_{g' \to h} \leqslant t_{end}|\alpha_g, \beta_g)(1 - \prod_{p'}(1 - P(t_{g' \to h} < t_{g'} + t_{p' \to h}|\alpha_g, \beta_g, \alpha_p, \beta_p))) \end{aligned} \tag{A.4}$$

Base rate activations of the effect event are represented as having been caused by the previous base rate activation, which can also be represented as $g' \to e'$ pairs where g' is actually the target component's (i.e. E) activation. When there are presumed preventative cause events, the base rate activation could be prevented but then subsequently "recover". Therefore, for base rate activation we could jointly consider Step 1 and Step 2 as $g' \to h^{(1)} \to \ldots \to h^{(n)} \to e'$, where $h^{(1)} \ldots h^{(n)}$ happens within the preventative windows. Meanwhile, according to the transition property of the gamma distribution, if $X, Y \sim Gamma(\alpha, \beta)$ then $X + Y \sim Gamma(2\alpha, \beta)$. The probability $P(g' \to h^{(1)} \to \ldots \to h^{(n)} \to e')$ can thus be represented as Eq. A.5, where the calculation of $P(g' \to e')$ is similar to Eq. A.3, and the calculation of $P(g' \to h^{(n')})$ is similar to Eq. A.4 except that t_{end} is substituted with $t_{e'}$ and only the second item of prevention is considered.

$$\begin{aligned} P(g' \to h^{(1)} \to \ldots \to h^{(n)} \to e'|\alpha_b, \beta_b, \alpha_p, \beta_p) = \\ P(g' \to e'|(n+1)\alpha_b, \beta_b) \prod_{n' \in n} P(g' \to h^{(n')}|n\alpha_b, \beta_b, \alpha_p, \beta_p) \end{aligned} \tag{A.5}$$

Finally, the prevention examination in Step 3 extracts all presumed preventative events and their nearest effect events to form $p' \to e'$ pairs (there is no need for examination if no effect events happen after p'), and then applies the gamma cumulative density function of prevention:

$$P(p' \to e'|\alpha_p, \beta_p) = P(t_{p' \to e'} < t_{p'e'}|\alpha_p, \beta_p) \tag{A.6}$$

A.2 Implementation of simulation-and-summary-statistic models

A.2.1 Cue distributions

We constructed the cue distributions (see Figure 5.3b) for each type of connection (generative, non-causal, preventative) under two base rates (regular, irregular) by simulating 90,000 interactions with imagined causal devices. These included 10,000 simulations of each of the 9 causal structures considered here. In each simulation the structure is perturbed by interventions performed in random orders with random timings.[1] In this way we establish a marginal distribution for each summary statistic under each type of connection. Note that we used a large number of simulations to produce smooth distributions for our later model fitting, however similar distributions can be achieved with a much smaller number of simulations (Ullman et al., 2018). As shown in Figure 5.3b, the delay cue is independent of questions of segmentation by definition since it always relates to the earliest subsequent effect event after each intervention. The count cue, however, is sensitive to the choice of segmentation, meaning we consider intervention-window and fixed-window assumptions separately. For delay distributions, we use a probability density function smoothed with Gaussian kernels, while for count distributions we can use the discrete probability mass functions directly.

A.2.2 Likelihood calculation

We assume each connection is estimated independently as either generative, non-causal, or preventative, and then combined to yield an overall probability for each candidate causal structure. For example, an intervention on A with the nearest effect occurring 2.5 seconds later has a likelihood of [.2, .7, .1] of having been produced by a generative, non-causal or preventative $A \rightarrow E$ connection respectively under the regular base rate and [.3, .6, .2] under the irregular base rate. When the next intervention on A happens, the posterior is updated by taking the product of this new likelihood with the preceding ones.

A.2.3 Boundary situations

We consider boundary situations when observing evidence as follows: If no effect occurs within the observation window, in both segmentation approaches, the delay cue will be marked as larger than the observation window and the probability is estimated according to the cumulative density function falling after this. If the observation window is less than the fixed window length for

[1] Similar to generating the experimental stimuli, each simulation included three interventions on A and three interventions on B. Distinct from the experimental stimuli, simulated sequences here were not cut at twenty seconds so as to avoid the complex boundary effects in distribution construction.

the fixed-window approach (which often happens near the end of the clip), or there is no next intervention in the intervention-window approach, the count cue will be marked as greater than or equal to the observed count of effects and the probability is also estimated on the basis of its cumulative mass function.

A.3 Model fitting procedure

We considered four models in total:

1. Fully normative inference based on marginalizing over all possible causal pathways.
2. Summary-statistic (SS) based inference, using a fixed 4 second window to count events following each intervention.
3. Summary-statistic based inference, using the interval until the next intervention to count events.
4. A parameter free baseline that predicts all structure judgments to be selected with equal probability.

As in our comparison to simulations, we simply assume the delay and count cues are equally weighted and merged. We assume learners begin each problem with a uniform prior over causal structures. We feel this is a reasonable choice here since the relatively small hypothesis space,a balanced set of trials, and the abstract setting leave little for inductive biases to attach to. Nevertheless, we accept that we cannot rule out the possibility that some of the findings we attribute to evidence processing enter through prior preferences. To map models' posterior probabilities to judgments, we assumed participants' responses result from a softmax over a posterior probability vector v:

$$P(n) = \frac{\exp(v_n/\tau)}{\sum_{n' \in N} \exp(v_{n'}/\tau)} \tag{A.7}$$

The "temperature" parameter $\tau \in (0, +\infty]$ controls how reliably the participant selects the most probable answer (i.e. that with the largest v_n in choice n). Smaller τ connotes higher choice reliability with $\tau = 0$ corresponding to hard maximization and $\tau \to \infty$ approaching random responding.

We evaluate model fit using cross-validation. At the aggregate level, we fit parameters to the judgments from $K - 1$ subsets of the complete dataset, and evaluate model performance in terms of its log-likelihood of predicting the left-out subset. K was defined via the stimulus seeds in each experiment (i.e. $K = 18$ in Experiment 1a and $K = 12$ in Experiment 1b including stimuli with and without a ground truth). This provides a rigorous and generalizable test of the models, since the actual sampled values of the stimuli (e.g. intervention timing, base rate activating timing,

etc.) are always outside of the training sample for all test sets. On the individual level, we similarly applied hold-one-stimulus-out as our cross-validation scheme for all experiments. For easy familiarity and comparability with other model based analyses of causal learning data, we also report Bayesian Information Criterion (BIC) penalized fits to the full dataset.

A.4 Alternative model fitting results

Table A.1: Model fits separated by conditions.

	Regular			Irregular		
	CV	BIC	τ	CV	BIC	τ
Experiment 1a						
Normative	-2894	5789	0.45	-3162	6327	0.43
SS (intervention-window)	**-2822**	**5648**	0.23	**-3036**	**6076**	0.22
SS (fixed-window)	-2924	5853	0.31	-3074	6153	0.29
Random	-3695	7390		-3735	7469	
Experiment 1b						
Normative	-2256	4497	0.62	-2167	4332	0.51
SS (intervention-window)	**-2041**	**4086**	0.23	**-2014**	**4032**	0.24
SS (fixed-window)	-2114	4232	0.35	-2052	4106	0.31
Random	-2503	5006		-2384	4768	

Table A.2: Model fits separated by blocks in Experiment 1b.

	Ground Truth			No Ground Truth		
	CV	BIC	τ	CV	BIC	τ
Normative	-2009	4022	0.46	-2361	4725	0.88
SS (intervention-window)	**-1917**	**3840**	0.24	**-2141**	**4279**	0.23
SS (fixed-window)	-1982	3969	0.32	-2188	4373	0.35
Random	-2443	4887		-2443	4887	

Table A.3: Model fits with one cue in summary-statistic models.

	Delay			Count		
	CV	BIC	τ	CV	BIC	τ
Experiment 1a						
SS (intervention-window)	-5994	11990	0.31	-6065	12134	0.20
SS (fixed-window)	-6040	12084	0.35	-6228	12460	0.33
Experiment 1b						
SS (intervention-window)	-4136	8277	0.33	-4173	8343	0.20
SS (fixed-window)	-4196	8393	0.39	-4292	8585	0.36

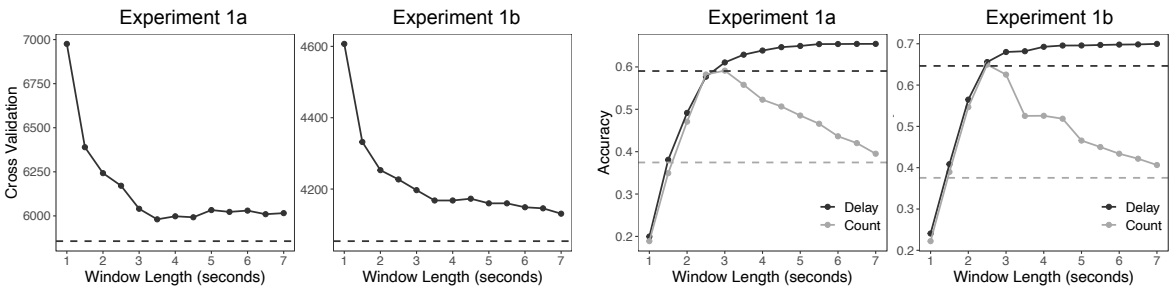


Figure A.1: Cross validation results and model accuracy under different fixed-window lengths for summary-statistic models. Horizontal dashed lines indicate cases of intervention-window segmentation.

Appendix B

Appendices for Chapter 6

B.1 Ideal observational (IO) learning

We formulate how a learner should ideally update their beliefs after seeing evidence produced by interventions. The ideal observer infers a posterior distribution $P(S|\mathbf{d};\mathbf{i})$ over causal structures $s \in S$ based on evidence $\mathbf{d}$ conditional on interventions $\mathbf{i}$ using the Bayes rule:

$$P(S|\mathbf{d};\mathbf{i}) \propto p(\mathbf{d}|S;\mathbf{i}) \cdot P(S). \tag{B.1}$$

Here, $P(S)$ denotes the prior probability distribution over causal structures, and $p(\mathbf{d}|S;\mathbf{i})$ denotes the likelihood of the observed data conditional on the interventions under each possible causal structure.

We assume that data $\mathbf{d}$ consists of all non-interventional activation events and their timings indexed by their chronological order and subscripted by the component at which they occur $d^{(\text{index})}_{\text{component}}$ and that this is conditioned on the set of interventions $\mathbf{i}$ including all activations $a^{(\text{index})}_{\text{component}}$ and blocks $b^{(\text{index})}_{\text{component}}$ performed by the learner during the learning episode.

As mentioned in the main text, when calculating the likelihood of the data given a candidate structure, there are likely to be multiple potential paths of actual causation. Each of these has its own likelihood. To construct the total likelihood of a hypothesized causal structure and interventions producing a set of events, we must consider all possible causal paths $\mathbf{Z}_s$ that could describe what actually happened given structure s and then repeat this for every $s \in \mathbf{S}$. Since the path set is exclusive and exhaustive conditional on the structure under consideration s, we can sum the path likelihoods to calculate the total likelihood of that structure producing the data:

$$p(\mathbf{d}|s;\mathbf{i}) = \sum_{z' \in \mathbf{Z}_s} p(\mathbf{d}|z';\mathbf{i}). \tag{B.2}$$

To construct the possible paths, each effect event must be attributed to exactly one preceding event occurring at a component with a causal link to that effect in structure s. Assessing the likelihood of each valid path includes two parts: (1) Explaining all actual effects; (2) explaining away any expected effects that did not occur. The first part is just the product of the gamma densities for all the causal delays between observed effects and their putative causes. Each delay is given by

$$\gamma_{pdf}(t_{\text{effect}} - t_{\text{cause}}, \alpha, \beta). \tag{B.3}$$

For the latter part, we need to check that each cause event assumed by the hypothetical structure has its corresponding effect(s) in the path. Each one that is missing must have failed (with probability $1 - w$), or (with probability w) be either yet to occur or have been blocked from occurring. Combining these possibilities we get the following expression:[1]

$$w\Big[\underbrace{\gamma_{cdf}(\max(0, t_{\text{block}_{\text{onset}}} - t_{\text{cause}}), \min(t_{\text{block}_{\text{offset}}} - t_{\text{cause}}, t_{\text{now}} - t_{\text{cause}}), \alpha, \beta)}_{\text{activation was blocked}} + \tag{B.4}$$

$$\underbrace{(1 - \gamma_{cdf}(t_{\text{now}} - t_{\text{cause}}, \alpha, \beta))}_{\text{activation has not occurred yet}}\Big] + \quad \underbrace{(1 - w)}_{\text{activation failed}} \tag{B.5}$$

Thus, the likelihood for a particular causal path given a particular causal structure can be calculated exactly via a combination of diagnostic reasoning — attributing exactly one cause for each observed effect — and predictive reasoning — attributing exactly one effect or failure to each causal link coming out of each activated component.

B.2 Comparing simulated resource-rational interventions and judgments

To test whether our intervention and judgment models can replicate participants' qualitative behavior patterns, we use the parameters fit by human data to simulate resource-rational agents that intervene on the same devices examined in Experiments 1 and 2. In both the reliable and unreliable delay conditions, we generated 30 simulated learners. The intervention patterns are shown in Figure B.1. This shows that for both experiments, simulated resource-rational learners activated components in acyclic structures more than cyclic structures, and four-node structures more than three-node structures. They waited longer to perform their next intervention when the structures were cyclic then when they were acyclic. For Experiment 2, they performed more blocking actions in cyclic devices than acyclic devices. These results demonstrate that our intervention model is capable of replicating a wide range of human intervention patterns.

[1] $\gamma_{cdf}(x, y, \alpha, \beta)$ in Equation B.5 denotes the cumulative probability of a delay being between x and y in length.

We also provided simulated evidence to the judgment model, which was based on parameters fitted with human data. For both experiments, the noisy-IO judgment model replicated the human result that acyclic structures had higher accuracy than cyclic structures. Unlike participants, these simulations were not more accurate on unlinked structures in Experiment 2. This could be due to some extra assumptions that we did not include in our models, such as the possibility that rather than beginning each trial with a uniform prior over structures, participants may have expected causal models to be sparse (Lu et al., 2008).

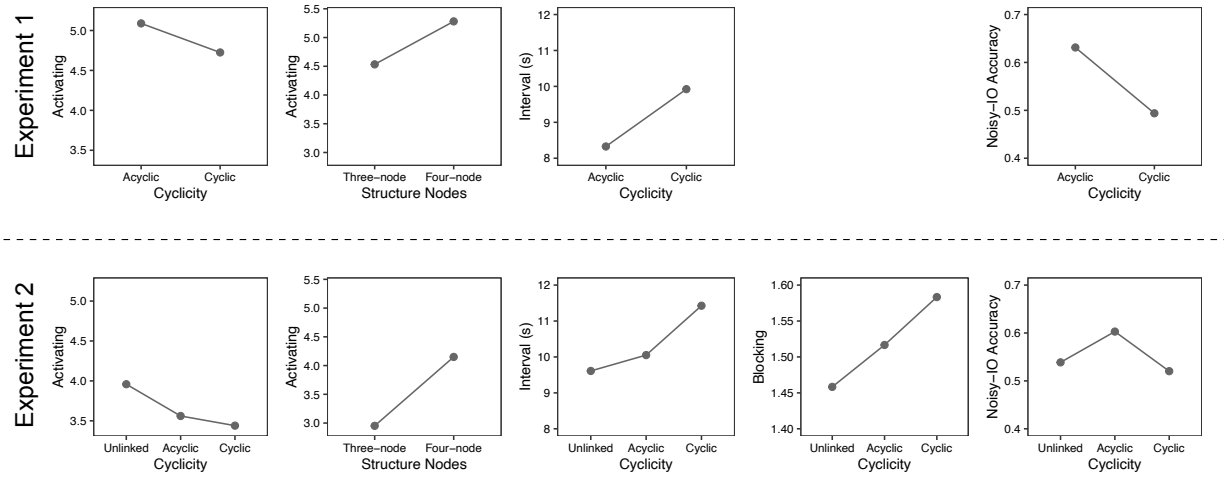


Figure B.1: Results from simulated evidence according to the parameters fit in the intervention and judgment models.

B.3 Supplementary tables

Table B.1: Intervention model fits of back tracking window size in polynomial local cost models with a generic exponent of 2.

Window	CV	BIC	τ	ω	θ
Experiment 1					
1 s	-16426	32810	2.21	5.38×10^{-2}	7.77
2 s	-16430	32797	2.23	1.94×10^{-2}	7.82
3 s	-16414	32789	2.24	1.17×10^{-2}	7.82
5 s	-16418	32798	2.30	7.30×10^{-3}	7.99
6 s	-16421	32805	2.34	6.46×10^{-3}	8.09
7 s	-16425	32814	2.38	6.12×10^{-3}	8.21
Experiment 2					
1 s	-40466	80811	1.86	8.81×10^{-3}	7.66
2 s	-40459	80798	1.86	3.42×10^{-3}	7.63
3 s	-40457	80794	1.85	2.17×10^{-3}	7.60
5 s	-40467	80813	1.86	1.29×10^{-3}	7.65
6 s	-40471	80821	1.87	1.11×10^{-3}	7.67
7 s	-40477	80833	1.88	9.74×10^{-4}	7.72

Table B.2: Intervention model fits of exponents or base parameters in polynomial- or exponential-costs.

Experiment 1						
	CV	BIC	$c_{CV/BIC}$	τ	ω	θ
Global cost						
EIG-ECC$_{\text{Linear}}$	-16550	33040		2.00	1.00×10^{-1}	6.80
EIG-ECC$_{\text{Polynomial}}$	-16451	32882	1.8	2.55	1.03×10^{-2}	8.52
EIG-ECC$_{\text{Exponential}}$	-16550	33067	2.4/1.8	2.79	1.44×10^{-10}	10.00
Local cost						
EIG-ECC$_{\text{Linear}}$	-16524	32994		1.82	1.32×10^{-1}	6.13
EIG-ECC$_{\text{Polynomial}}$	-16378	32743	1.6/1.4	1.80	6.22×10^{-2}	6.04
EIG-ECC$_{\text{Exponential}}$	-16556	33083	1.8	2.79	1.39×10^{-10}	10.02
Experiment 2						
	CV	BIC	$c_{CV/BIC}$	τ	ω	θ
Global cost						
EIG-ECC$_{\text{Linear}}$	-40507	80898		1.98	3.33×10^{-6}	8.13
EIG-ECC$_{\text{Polynomial}}$	-40504	80901	3	1.96	4.40×10^{-6}	8.06
EIG-ECC$_{\text{Exponential}}$	-40503	80895	1.2/1.18	1.97	7.31×10^{-11}	8.09
Local cost						
EIG-ECC$_{\text{Linear}}$	-40507	80898		1.98	3.76×10^{-6}	8.13
EIG-ECC$_{\text{Polynomial}}$	-40462	80815	2	1.86	1.59×10^{-3}	7.63
EIG-ECC$_{\text{Exponential}}$	-40500	80896	1.18/3.2	1.97	8.65×10^{-11}	8.10

Note: The base and exponent parameter c was fit by grid search. We searched in (1,4), using a step of 0.002 for the range [1.002,1.018], a step of 0.02 for the range [1.02,1.18] and a step size of 0.2 thereafter. These steps approximate log-uniform intervals so are suitable for fitting a parameter bounded at the lower end but not at the higher end. Other parameters were fitted given a fixed c. We reported two c with CV and BIC deviated in their results.

Table B.3: Retrospective local complexity model fits.

	CV	BIC	τ	ω	θ
Experiment 1					
EIG-ECC$_{\text{Linear}}$	-16566	33077	3.11	3.30×10^{-1}	10.83
EIG-ECC$_{\text{Polynomial}}$	-16588	32120	3.01	1.67×10^{-5}	10.80
EIG-ECC$_{\text{Exponential}}$	-16620	33120	3.00	1.57×10^{-7}	10.80
Experiment 2					
EIG-ECC$_{\text{Linear}}$	-40507	80898	1.98	4.61×10^{-5}	8.13
EIG-ECC$_{\text{Polynomial}}$	-40506	80898	1.98	4.74×10^{-6}	8.12
EIG-ECC$_{\text{Exponential}}$	-40507	80898	1.98	3.01×10^{-11}	8.13

Note: To fit retrospective models, we replaced the complexity component in Equation 6.7 with a retrospective recent event count. This assigned a cost according to the number of events in previous 4 seconds for activation interventions $\{a_A, a_B, a_C, a_D\}$, while assigned a cost of zero for doing nothing or blocking interventions.

Appendix C

A supplementary experiment for Chapter 7

The Trend × Instruction interaction in Experiment 1 in chapter 7 showed that people reacted to instructions to use different perspectives of temporal information to make causal inferences. The supplementary experiment aimed to investigate whether or not the effect came from the instruction manipulation or the simple format difference. In this experiment, we simply manipulated the display format to either imply a complete or an ongoing set of measurements, without the different text prompts.

C.1 Method

Participants Two-hundred participants (120 female, 78 male, 1 non-binary, 1 unenclosed, aged 44 ± 13) were recruited from Prolific Academic and were randomly assigned to either the Closed (N=100) or Open (N=100) conditions (see Design & Materials below).

Design & Materials The experimental design and materials were very similar Experiment 1, except that now both groups were exposed to a neutral instruction "The observation has happened for five days so far. The records now include Day 1 to Day 5". We hence here renamed the between-subject manipulation as a pure "Format" (Closed vs. Open) manipulation.

C.2 Results

Similar to Experiment 1, there is a main effect of contingency ($F(2, 198) = 170.86$, $p < .001$, partial $\eta^2 = .46$; pairwise comparison: zero–beneficial: $t(198) = 12.11$, $p < .001$, $d = 0.58$; zero–harmful: $t(198) = 8.58$, $p < .001$, $d = 0.36$; harmful–beneficial: $t(198) = 15.13$, $p < .001$, $d = 0.94$

under Bonferroni's adjustment, Figure C.1). There was no main effect of Trend ($F(1, 198) = 0.08$, $p = .77$) or Instruction ($F(1, 198) = 0.19$, $p = .66$). In contrast to Experiment 1, there were no any two or three-way interaction effects ($ps > .05$, Figure C.2). This shows that the influence of Instruction in Experiment 1 cannot be simply replaced with format differences.

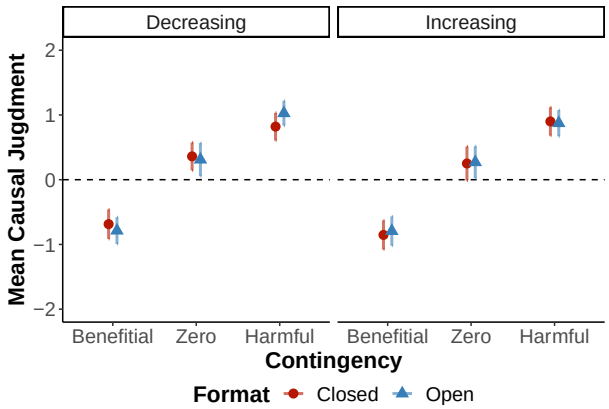


Figure C.1: Causal judgments under different contingency and experimental conditions. Participants judged the influence of treatment on a scale from -3 (definitely beneficial) to 3 (definitely harmful). Higher scores mean people are more sure the treatment is harmful to the bacteria survival while lower scores mean people are more sure the treatment is beneficial. 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.

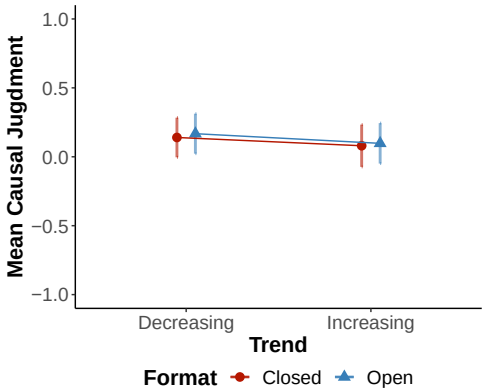


Figure C.2: Causal judgments under Decreasing vs. Increasing trends across experimental conditions. Participants judged the influence of treatment on a scale from -3 (definitely beneficial) to 3 (definitely harmful). Error bars indicate 95% confidence intervals.