1

Unifying Intervention and Time in Causal Learning

Neil R. Bramley (NYU), Tobias Gerstenberg (MIT), Ralf Mayrhofer (Göttingen), and David A. Lagnado (UCL)

Abstract

Much of what we know about the world comes from acting on it, and observing the consequences of our actions. In the literature, causal learning from interventions and from observing temporal dynamics have largely received separate attention due to the different datasets they are usually applied to. However, we argue that in human cognition, interventions and temporal dynamics are inseparable. We trace how causal inference tools developed in data science have been applied to understanding human causal learning and reasoning, highlight the current shortcomings of both intervention-based and time-based approaches taken separately, and describe recent work that starts to bring the two together. We end by sketching an account of interventional and temporal evidence as constituents of a unified online causal learning process.

not sure about the title of the paper; it sounds a little strange to unify time and intervention since these are conceptually different things

1.1 Introduction

Uncovering and describing the deep causal structure of reality is a fundamental goal of science, but it is also at the heart of human cognition. We are born into a "blooming, buzzing confusion" (James, 1890)

need to add page number of direct quote

of sensory information, and spend much of our development building up a causal model of reality that is both rich enough and accurate enough to guide us in pursuing our changing goals. On this view, sci-

ence plays a supporting role in causal cognition: extending the domain of causal understanding beyond what can be inferred through everyday experience; integrating evidence at scales beyond the capabilities of the brain; and so helping resolve disagreements about hypothesized causal connections. However, attempts to understand the psychology of causal leaning and reasoning have recently flowed in the other direction.

wasn't obvious to me what was meant by 'other direction' here

Bayesian networks have become a ubiquitous tool for modeling both non-causal and causal inference in large data sets across the sciences (Pearl, 1988). Among other things, they provide a convenient calculus for learning from, and reasoning about, *interventions* — actions or experiments that manipulate variables in the world (Pearl, 2000; Woodward, 2003).

what are 'variables in the world'? they exist in a graph but not really in the world

As a result they have also been applied to the task of modeling human active causal learning and reasoning. Separately, a plethora of statistical methods are used for drawing causal inferences from time series data (Friston et al., 2014; Granger, 2004) and a separate line of research explores the interplay between time, causal beliefs, and predictive perception (e.g., Bechlivanidis & Lagnado, 2016; Buehner & McGregor, 2006).

In this chapter, we contend that the division in the scientific literature between intervention-driven and time-driven modes of causal inference is an artifact of the kinds of datasets that scientists have to work with. Experimental datasets are typically aggregated over many independent samples from a population in a setting where temporal dynamics are unavailable or unmeasured, while nonexperimental datasets often track multiple factors across time. We argue that there is no such data distinction in human experience — we experience the world as a single ongoing event stream and are constantly choosing how and when to next intervene.

this might be a little strong – it's probably true about the experience of directly interacting with the world, but we also learn from data that takes more abstract forms; so the statement is probably only true about learning before cultural artifacts (also ignoring social learning of course)

We must be sensitive what goes on before, during, and after our actions if we want them to be effective or informative. As such, we argue that we should move toward modelling

make sure to use american or british spelling consistently

human causal learning as an inherently online process, involving the continuous integration of both temporal and interventional information. We illustrate some shortcomings of treating interventions and time as separate cues through reference to past research, and then highlight new approaches and data that pave the way for a new conception of the role of intervention and time in human causal learning.

we don't want to talke prior research down; rather spin it positively: how we are building on it

maybe we should have a little more detailed roadmap for the rest of the paper here?

1.2 The Causal Bayesian network account of learning from interventions

A core challenge for causal inference both in science and in everyday human learning is distinguishing genuine causal relationships from spurious or coincidental ones. Many things in the world are statistically associated — sun exposure is associated with melanin production; smoking is associated with lung cancer; ice cream sales are associated with deaths by drowning; and homeopathic remedies are often associated with positive health outcomes. Some of these associations are due to direct causal relationships — sun exposure really causes skin to tan and smoking really causes cancer. But, in many others cases, the relationship is due to shared causal factors — people are both more likely to eat ice cream and go swimming when the weather is warm, and people often feel better after taking a medicine they believe works, regardless of whether it is actually effective (Di Blasi, Harkness, Ernst, Georgiou, & Kleijnen, 2001). Intervention is a way of assessing whether an association is really causal. Interventions are actions that perturb the world and so reveal its structure in ways that are unavailable from mere observation (Woodward, 2003). In science we call these experiments, and plan them carefully, systematically manipulating things on a large enough scale to resolve causal questions of interest in the face of irreducible noise. When we manipulate one variable we no longer have to worry that a resulting association between that variable and another is due to a shared cause. For instance, an experiment systematically exposing some participants to sunlight and others to darkness, measuring their skin tones before and

after, will reveal a positive relationship between the intervened-on factor (sun exposure) and its putative effect (skin tone). In contrast, systematically giving some people homeopathic remedies and others sugar pills without telling them which, normally removes any relationship between treatment (whether the pill is a homeopathic remedy) and its putative effect (health outcomes), revealing that it is the belief in the efficacy of the drug rather than the drug itself that causes the association.

i'd remove the last paragraph about homeopathy

Interventions in human experience are more diverse and ubiquitous than those practiced in science. Our every action affects our proximal world in some way, from small movements that affect our pose or field of view, to the extended actions through which we interact with other objects and one another. Sometimes we act with the goal of resolving some particular causal uncertainty ("What does this button do?", "What does this taste like?","What will happen if I shout loudly in the library?"), other times we act primarily to pursue our goals (turning on the PC, feeding ourselves, getting someone's attention). Causal Bayesian networks (Pearl, 2000) provide a framework for drawing inferences based on both observations and interventions. We now briefly describe this framework and how it has been applied to the study of causal cognition.

1.2.1 The causal Bayesian network framework

Causal Bayesian networks (hereafter "CBNs", Pearl, 2000) capture aggregate patterns of covariation between variables in terms of probabilistic causal dependencies.

first sentence is most likely confusing to a less informed reader – maybe just desribe what they are first, before saying what they do

Nodes represent variables (i.e. the component parts of a causal system); arrows represent causal connections; and parameters encode the combined influence of parents (the source of an arrow) on children (the arrow's target, see Figure 1.1a for an example). CBNs can represent discrete or continuous valued variables and the functional dependence between the state of an effect on the states of its causes can take arbitrary form. However, the majority of psychology research has focused on systems of binary $\{0 = \text{absent}, 1 = \text{present}\}$ variables and has commonly assumed a simple parameterization for both generative and preventative causal relationships that we describe in more detail below (Cheng, 1997).

Figure 1.1a depicts a causal network relating three binary variables

X_1, X_2 and X_3, with one generative $X_1 \xrightarrow{+} X_2$ connection and one preventative $X_2 \xrightarrow{-} X_3$ connection. Under this model, X_1, X_2 and X_3 all occur with their own "base rate" probability (i.e. due to causes exogneous to the model). However the probability that X_2 and X_3 occur also depend on the state of their causes, such that $P(X_2 = 1)$ is higher than baseline when X_1 is present while $P(X_3 = 1)$ is lower than baseline when X_2 is present.

Figure 1.1b depicts some possible observations $\mathbf{d} = \{d_1 \dots d_8\}$ produced by the causal system in Figure 1.1a. Any parameterized causal model s over variables $\mathbf{X} = \{X_1 \dots X_N\}$ assigns a probability to each datum $d = \{X_1 = x_1 \dots X_N = x_n\}$, meaning that Bayesian inference can be used to assess which of a set of potential CBNs (which $s \in \mathcal{S}$) does the best job of accounting for data. This will be whichever model(s) capture the pattern of statistical (in)dependencies between the variables with the minimum number of connections.

Bayesian networks embody the ambiguity described above, about how observed correlations relate to causality. Without causal insight, we might construe the dependence between X_1 and X_2 and between X_2 and X_3 in several different ways. One is as depicted ($X_1 \xrightarrow{+} X_2 \xrightarrow{-} X_3$), but the reverse ($X_1 \xleftarrow{+} X_2 \xleftarrow{-} X_3$) is also possible, as is the case where where both X_1 and X_3 depend on X_2 (i.e., $X_1 \xleftarrow{+} X_2 \xrightarrow{-} X_3$). In each case, the best fitting parameters $\mathbf{w}_s$ differ, but the marginal and conditional probability structure is preserved and the overall goodness of fit to the data will be the same. This property of Bayesian networks is known as the Markov condition (Pearl, 2000).

Adding the notion of an intervention breaks this deadlock. Metaphysically, an intervention in a CBN is conceived as the learner first setting the values of some of the variables, and then allowing the causal system to propagate this through the rest of the network.

maybe state this more generically – it's not necessarily about a 'learner'

This is analogous to reaching into the causal system, changing things within it, then observing what happened as a result. We model interventions in a CBN by fixing "intervened on" variables to their chosen values and disconnecting them from their normal causes, using Pearl's do[.] operator (Pearl, 2000) to denote what is fixed on a given test. Figure 1.1c gives an example in which X_2 is intervened on and set to 1 (i.e., do$[X_2 = 1]$). If the model is correct, we should expect X_1 to be present with its base rate probability, while X_3 should be subject to the preventative influence of X_2 (i.e. it should be less likely to occur than

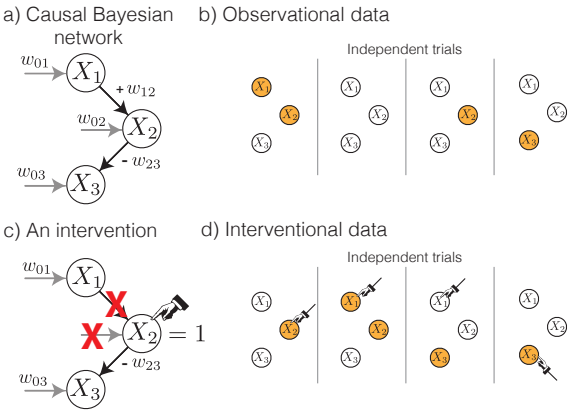

Figure 1.1 a) A causal Bayesian network containing a generative connection ("+" symbol) and a preventative connection ("-" symbol) parameterized with base rates w_{0i} and causal strengths w_{ji}. b) Example observational data. Yellow shading indicates a variable was present (i.e., took the value 1) while indicates a variable was absent (i.e., took the value 0) c) An intervention do$[X_2 = 1]$, disconnecting X_2 from X_1 and any background factors. d) Example interventional evidence.

if X_2 had been set to 0). Figure 1.1d gives examples of interventional evidence, under which various subsets of the variables are set through intervention either to 1 or 0 with the remaining variable states generated by running the causal system.

for figure d) it might be more straightforward to just show interventional data for the intervention shown in c); that's at least the default way in which a reader is likely to initially interpret the figure (since that pattern holds for a) and b)

1.2.2 Modeling intervention choice

Different interventions yield different outcomes, which in turn have different probabilities under different models. This means that which interventions are valuable for identifying the true structure depends strongly on the hypothesis space and prior. For instance fixing X_2 to 1 (do$[X_2 = 1]$) is (probabilistically) diagnostic if you are primarily unsure whether X_2 causes X_3 because $P(X_3|\,\text{do}[X_2 = 1])$ differs depending whether X_2 causes X_3. However, it is not diagnostic if you are primarily unsure whether X_1 causes X_2 because X_2 will take the value 1 irrespective of the causes of X_2. Optimal experimental design theory allows us to reason about the expected value of different interventions relative to a notion of

uncertainty (Fedorov, 1972; Raiffa, 1974). For instance, we can define the value of an intervention as the expected reduction in uncertainty about the true model after seeing what happened. This expectation can be calculated by averaging, prospectively, over the different possible outcomes of each potential intervention and computing the reduction in uncertainty in each case (e.g., Shannon, 1951). See Bramley, Dayan, Griffiths, and Lagnado (2017) for a detailed mathematical account of this.

A key property illustrated by Figure 1.1 is that well-chosen interventions can be much more informative, on average, than mere observations of the causal system. The depicted observational data (Figure 1.1b) is much less informative than the interventional data (Figure 1.1d) with respect to the set $\mathcal{S}$ of all possible causal Bayesian networks relating the three variables. We demonstrate this in the Figure

which figure?

by computing the posterior Shannon entropy $\mathrm{H}(P(\mathcal{S}|\mathbf{d};\mathbf{a}))$ after each observation or intervention, based on an initially uniform prior over structure hypotheses $P(\mathcal{S})$ and model parameters $p(\mathbf{w}_s)$, including these calculations in the Appendix.

If we think we need this...? There are lots of mathematical details commented out here that could be repatriated to an appendix.

personally, i don't think we should have an appendix. i'd say we just don't want any mathematical details – it's fine to just refer to some of your work here like you've done before

This is partly because the observational data does not distinguish between models with the same dependency structure

commented out these structures as you had mentioned them before

, but also because a well-selected sequence of interventions can systematically target and resolve residual sources of uncertainty, for instance by testing a particular causal connection that previous tests have not provided clear evidence about.

this section feels like it's part of causal cognition already (since it's about a learner trying to minimize uncertainty – so the next subheading is a little surprising

1.2.3 Interventions in causal cognition

CBNs were initially adopted in psychology because of their ability to account for qualitative patterns of human judgments that are hard to

capture under simple associative accounts of learning (e.g., Rescorla & Wagner, 1972). For example, CBNs capture "explaining away" effects where the causal roles of variables lead to asymmetries in judgments of their associative strength — competing causes share associative strength while effects do not (Holyoak & Cheng, 2011).

i don't think the 'explaining away' will be understood as currently stated

They also capture "blocking" effects in which learning the state of a mediator or common-cause overrides inference from one of the distal variables to the other.

this sentence is also a little too abstract – maybe start with the concrete example and then mention the more general principle; also "blocking" is only informative for someone who already knows the associative learning terminology – many philosophers won't

For example, in Figure 1.1, knowing that $X_1 = 1$ only tells you something about $P(X_3)$ (i.e., that it is less likely) if you do not already know the value of X_2. More generally, CBNs have shed light on how people go from from contingency information to judgments of causal strength through the idea that people interpret evidence through the lens of their favored causal model, meaning background factors and surrounding causes affect the interpretation of evidence (Cheng, 1997; Griffiths & Tenenbaum, 2005; Waldmann, 2000).

last sentens is too long and has to manny "through"s; i would shorten this previous paragraph

Pearl's "Do" calculus has also proven to be an effective way of capturing the different inferences people make from observations and interventions (Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005), and how they reason about counterfactuals (Lagnado, Gerstenberg, & Zultan, 2013; Rips, 2010; Rips & Edwards, 2013).

you might need to say a little more about what counterfactuals are if you mention it here

The broad idea is that observations license backward inferences while interventions do not. Observing that your officemate arrives at the office soaked suggests may be raining outside, while if you had intervened and poured a bucket of water on them, their wet clothes would not tell you anything about the weather. The notion that people can imagine virtual interventions helps explain important aspects of thinking. For example, counterfactual "what if..." inferences are often consistent with the idea that people imagine an intervention that makes the counterfactual true

with minimal revision of its causal history (Gerstenberg, Bechlivanidis, & Lagnado, 2013; Lagnado et al., 2013; Rips, 2010).

this section needs a bit of work i think – maybe best to just start off with an intuitive example that illustrates the difference between intervention and observation; say that this can be captured in a CBN but not within an associative framework, and then say something about interventions for hypothetical planning, and for explanations (via counterfactuals)

The importance of interventional learning is well established in education, and developmental psychology, where self-directed "play" is seen as vital to healthy development (e.g. Bruner, Jolly, & Sylva, 1976; Piaget & Valsiner, 1930). Accordingly, a number of developmental psychologists have adopted a "child as scientist" analogy, which views children as fundamentally engaged in causal hypothesis testing within the CBN framework (Gopnik et al., 2004; Gopnik & Sobel, 2000; Sobel & Kushnir, 2006). In adults, a number of studies have found that people benefit from the ability to perform (or watch others performing) interventions during causal learning (Lagnado & Sloman, 2002, 2004, 2006; L. E. Schulz, 2001; Sobel & Kushnir, 2006).

Several recent studies have also explored *how* people select what interventions to perform, comparing adults' and children's choices against the dictates of optimal intervention selection, considering constrained and heuristic variants of this (Bramley, Dayan, et al., 2017; Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015; McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). McCormack et al. (2016) had children learn about the deterministic causal relationships between three spinning wheels sticking out of a black box toy. They were allowed to optionally insert a "Stop" sign to block one of the wheels from turning (equivalent to fixing that variable to 0 i.e., *absent*), and would also choose a wheel to spin, so initiating activity in the system (equivalent to fixing a variable to 1, i.e., *present*) and would then see which non-fixed wheel(s) spun as a result. The children then selected which of several causal structure pictures described the internal mechanism. McCormack et al. (2016) found that the quality of children's intervention choices improved with age and was a determinant of judgment accuracy. However, children of all ages also had a strong tendency to intervene by turning on X_1 which was the root component of all three possible structures. This intervention was completely uninformative because it made all three variables spin whatever the structure (the possible mechanisms were $X_1 \xrightarrow{+} X_2 \xrightarrow{+} X_3$, $X_1 \xrightarrow{+} X_3 \xrightarrow{+} X_2$, or $X_2 \xleftarrow{+} X_1 \xrightarrow{+} X_3$).

Coenen et al. (2015), explored adults interventional learning — learners had to establish which of two wiring diagrams described a circuit boards by activating one of the components and seeing which others activated. They found that learners sometimes chose the most informative intervention, but as in McCormack et al. (2016), also frequently intervened on the component that made the most other components activate even when this was uninformative, doing so more often when under time pressure.

Bramley et al. (2015) and Bramley, Dayan, et al. (2017) explored adults' interventional causal learning in a series of tasks in which the target causal systems of interest were noisy CBNs with nodes visualized simply as abstract colored circles on a blank background that would change color when "activated" on a given trial (as in Figure 1.1c). Participants performed interventions by explicitly setting the states of any subset of these nodes and observing the results while making accuracy-incentivised judgments about the hidden causal relationships. Bramley et al.'s (2015) analysis showed that people's intervention choices were motivated by uncertainty reduction across the full space of possibilities but that learners' ability to integrate the evidence across trials was limited. Bramley, Dayan, et al. (2017) accounted for judgments and interventions pattern as resulting from a bounded inference process in which learners do not represent the full space of possibilities but rather make local changes to a single working hypothesis. However, a common judgment error in these studies was not captured by the bounded learning analysis. Participants still exhibited a tendency to mark direct connections from causes to both their direct and indirect effects, suggesting that people are hesitant about judging there to be indirect causation when shown multiple outcomes occurring simultaneously (see also Fernbach & Sloman, 2009; McCormack, Frosch, Patrick, & Lagnado, 2015).

The section above might be possible to cut substantially

1.2.4 Shortcomings of CBNs for modelling everyday interventions

In this section we highlight several mismatches between CBNs and the kinds of causal inference problems that characterize everyday causal cognition. These break down into: 1. The requirement that evidence be gathered over independent trials; 2. the absence of temporal considera-

tions relating interventions and subsequent measurements of the system; and 3. problems with representing cycles and feedback dynamics.

As we noted above, CBNs are very limited in their representation of time. The interventional calculus embodies the minimal assumption that causes precede their effects, but it says nothing about the relative delays of competing causal pathways. Thus, CBNs are a natural partner to evidence that really has been gathered across multiple independent trials, in which causal relationships play out too fast to measure, or in which each variable is only measured once. By design, these features are standard in real experimental data sets. For instance, medical studies will typically involve a procedure for randomized recruitment to ensure participants are roughly independent samples from the desired population. Then, there will be a protocol for when to measure outcome variables (such as blood pressure, insulin levels etc) following an intervention (i.e. delivery of a treatment). The use of a fixed trial protocol makes it straightforward to aggregate results across the sample of patients. However, choosing an appropriate schedule of treatment and measurement seems to require substantial preexisting causal expertise about how long the relevant mechanisms will take to work and when any effects will be most measurable.

A consequence of the CBN framework being widely used to study causal cognition is that many of the experiments in the literature provide learning data that is already "packaged" into a CBN suitable format. In early studies, training data was sometimes literally provided in tabular form, with participants were shown a table or icon array detailing how often several variables appeared together or separately and asked to make judgments directly from this (Cheng & Novick, 1990). In more recent studies, participants were shown a series of cases (Deverett & Kemp, 2012; Gopnik & Sobel, 2000; Lagnado & Sloman, 2002; Sobel, Tenenbaum, & Gopnik, 2004), or invited to choose a sequence of interventions of their own devising (Bramley, Dayan, et al., 2017; Bramley et al., 2015; Coenen et al., 2015). But in these studies participants are instructed or given a cover story implying that these should be treated as completely independent trials. In some cases, the cover stories involved artificial mechanisms that reset on each trial — this includes the "black box" toys used in developmental studies (Coenen, Bramley, Ruggeri, & Gureckis, 2017; Gopnik & Schulz, 2007; McCormack et al., 2016); but also a range of artificial mechanisms involving lights, sensors, circuits, and switches (Coenen et al., 2017; Sobel & Kushnir, 2006; Waldmann, 2000). In other cases, the nature of variables are not described at all

(Bramley, Dayan, et al., 2017; Bramley et al., 2015; Rehder & Burnett, 2005), or independence is established by instructing participants that each trial is performed with a different sample from a population (Rehder, 2003; Rehder & Waldmann, 2017; Rottman & Keil, 2012).

While ingenious, these cover stories create situations that rarely obtain in everyday learning. The causality we encounter in physical and social systems of everyday life do not, generally, reset themselves between each interaction. In life, one learning episode typically bleeds into the next, with no clear boundaries. We might try a medicine on multiple occasions, but if we want to treat these tests as independent we had better leave a substantial amount of time between each test, relative to our beliefs about how long the drug takes to pass through our system. We also must keep in mind the ongoing evolution of our health and potential adaptation in our receptivity to the medication. Choosing when to act, and how to delineate between and aggregate over "trials" in continuous experience are important aspects of real world causal reasoning (episodic cognition ref?), that are missed by the studies focusing on CBN-packaged data.

As well as separation *between* trials being hard to achieve in everyday life, sensitivity to how things play out *before, during* and *after* a real world interventions is also intuitively important. Rottman and Keil (2012) point out that it is often *change* in the state of the variable from one time point to another that people see as an effect needing an explanation. Furthermore, it is hard to imagine intervening on something in the world without first observing its pre-interventional state. The metaphysics of an intervention in a CBN not only presumes independence of one test from one another, but also imposes an implausible sequence of actions and measurements on the part of the learner. An idealized intervention involves setting variables substantially *before* observing the system. Any causal effects of one's interventions are assumed to have entirely propagated through the system by the time the observation is made, while the trace of the resulting variable states must persist long enough to be measured together at the end. To make this work in psychology experiments, cover stories are used that are generally either: vague about the measurement of time, pertain to variables whose states are only observable after the fact, or involve causal relationships that would propagate too fast to be observed. For example, Steyvers et al.'s (2003) cover story involved aliens reading one another's minds. It was implied that aliens would keep thinking the same symbol string (i.e. "XYZ") unless they chose to read another alien's mind, in which case

they would then think the same thought as that other alien and keep doing so until the measurement was made.

CBNs are based on factorization of a joint probability distribution, meaning they cannot naturally represent relationships that form loops or cycles. However, such dynamic relationships seem pervasive in the world — e.g. that liking rock music might make you join a rock band which might further increase your enjoyment of rock music and so on. All sorts of real world processes, from population change (Malthus, 1888) to economic, biological and physical interactions, are characterized by reciprocal and dynamical causal processes. In experiments, people frequently report causal beliefs that include cyclic relationships when allowed to do so (Kim & Ahn, 2002; Nikolic & Lagnado, 2015; Sloman, Love, & Ahn, 1998). While there are ways of adapting the CBN formalism to capture cycles — (e.g. Dean & Kanazawa, 1989; Lauritzen & Richardson, 2002) and see Rehder (2016) for a recent review — these either evoke a sequence of equally spaced discrete time steps or model the equilibrium behavior of the system. Thus, none of these proposals capture how cause–effect relationships unfold in continuous time, where some relationships might occur much faster or slower than others.

1.2.5 Summary and discussion

In sum, as we move away from carefully curated experimental scenarios toward more naturalistic interventional learning data, the assumptions that lend the CBN framework its mathematical simplicity also make it less adequate for the challenge. When we interact with the causal world, we typically have access to a lot more evidence than independent joint state measurements. We can monitor variables continuously, tracking when events occur relative to one another and how continuous quantities ebb and flow over time. We are sensitive not just to the "final" states of variables after causality has been and gone, but their states prior to interventions and their subsequent transitions. To intervene effectively and to draw sensible causal conclusions by aggregating over extended experience, we must not only worry about what variables are relevant, but also when they should be measured, how long to leave between tests, or how to best "reset" a system before testing it anew. This suggests that much of the most important and interesting causal inference work takes place while packaging a learning problem into a CBN suitable format. Independent trials are a luxury brought about through carefully curated scenarios, and aggregation to the level of contingencies and probabilities

only becomes possible when we have rich enough knowledge of functional form to abstract away from time's arrow. Thus, it is instructive to model this finer grain of human causal learning and reasoning in which time's arrow is fully represented. In the next sections, we build on these considerations and classic results about learning from temporal information, sketching two new approaches to modelling causal representation and interventional learning in continuous time.

1.3 Incorporating time: Event data

1.3.1 Existing research

People have been shown to make systematic use of both event order (Bramley, Gerstenberg, & Lagnado, 2014) and delay information in inferences about causal relationships (Buehner & May, 2003, 2004; Buehner & McGregor, 2006). Causal beliefs have also been shown to influence time perception (Bechlivanidis & Lagnado, 2013; Buehner & Humphreys, 2009; Haggard, Clark, & Kalogeras, 2002). A basic associative learning result is that, as the average interval between two events increases, the strength with which they are associated decreases (Grice, 1948; Shanks & Dickinson, 1987; Wolfe, 1921). However, people do not always see shorter intervals as more causal, but rather prefer intervals that match their expectations, where these might come from prior experience or though familiarity with causal mechanisms. Both shorter-than-expected and longer-than-expected intervals have been shown to reduce causal strength judgments (Buehner & May, 2002, 2003, 2004; Greville & Buehner, 2010, 2016; Hagmayer & Waldmann, 2002; Schlottmann, 1999). Reliability also appears to be key to strong causal attributions from time, with variability in inter-event intervals normally associated with reduced causal judgments (Greville & Buehner, 2010; Greville, Cassar, Johansen, & Buehner, 2013; Lagnado & Speekenbrink, 2010).[1]

Supporting the notion that temporal considerations inevitably feed into causal judgments, several studies have pitted temporal order cues against statistical contingencies. These studies generally found that causal judgments were dominated by temporal information (Burns & McCormack, 2009; Frosch, McCormack, Lagnado, & Burns, 2012; Lagnado & Sloman, 2004, 2006; Schlottmann, 1999). For example, Lagnado and Sloman (2006) explore a setting where a virus propagates through a network

[1] See Young and Nguyen (2009) for a counterexample.

of computers infecting each with some probability, but also with variable delays in transmission from one computer to another. Participants' task was to infer the structure of these computer networks based on having observed viruses spreading through the network on multiple trials. Participants preferred causal models that matched the experienced order in which computers displayed infection, even when covariational information (which subset of computers got the virus on each trial) suggested a different structure.

1.3.2 Representing causal events and interventions in time

To link these results to the problem of structure induction over multiple variables, we need a causal representation rich enough to encode beliefs about causal inter-event delays. Bramley, Gerstenberg, Mayrhofer, and Lagnado (2018) recently developed a normative framework that does this. Concretely, their framework captures causality between events (hereafter "activations") that occur at components of a system $X_1 \dots X_n$ at particular points in time, but have no measurable duration of their own (Cox & Isham, 1980). Bramley, Gerstenberg, Mayrhofer, and Lagnado's (2018) approach captures activation patterns within a causal system as being produced by mixture of exogenous influences and endogenous causal relationships with parametric delays governed by Gamma distributions with some mean μ and variability α. Figure 1.2a gives an example of an $X_1 \xrightarrow{+} X_2 \xrightarrow{-} X_3$ chain analogous to the CBN in Figure 1.1. As in a CBN we can assume all three variables are activated due to exogenous factors with their own "base rate" probability. This can be captured by a Gamma distribution where shape α is set to 1, equivalent to an Exponential distribution with rate $1/\mu$. This means that the expected delay before the next activation of X_1 is constant, regardless how long it has been since the last activation of X_1 or any other variable. Activations of X_1 then cause activations of X_2 with a typical delay, here captured by a Gamma($\mu = 1, \alpha = 10$) – i.e., with a mean of 1 second and standard deviation of 0.1 seconds (see Subfigure 1.2a, ii). These caused activations are in addition to any activations of X_2 caused by its base rate. Thus, the model implies that we should expect to see X_2 activate shortly after observing X_1 activating. Activations of X_2 then have a preventative effect on X_3. This is modeled by having activations of X_2 temporarily censor the next activation of X_3. Mathematically, this changes the distribution for X_3 from Exponential($\lambda = 1/y$) $\equiv$ Gamma($\mu = y, \alpha = 1$)

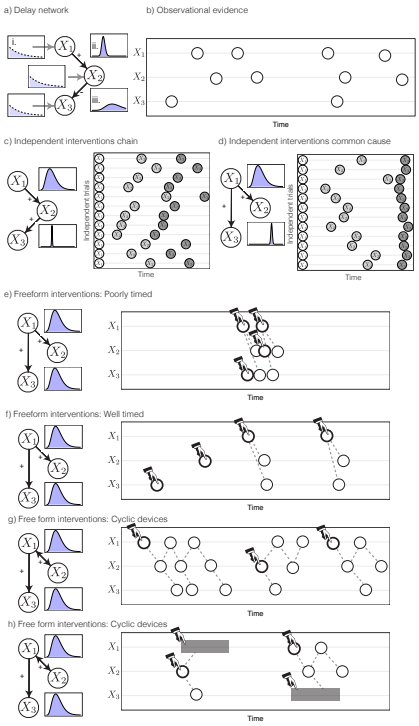

Figure 1.2 a) Example causal delay network with a generative $X_1 \xrightarrow{+} X_2$ preventative $X_2 \xrightarrow{-} X_3$ connection (see Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018; Bramley, Mayrhofer, et al., 2017). b) Example observational evidence. Rows denote variables and white circles mark activations over time. c) Independent interventions (rows) on a chain with unreliable $X_1 \xrightarrow{+} X_2$ connection and reliable $X_2 \xrightarrow{+} X_3$ connection. The timing of X_2 predicts the timing of X_3 because it is on the causal pathway. d) Same for common cause with reliable but longer $X_1 \xrightarrow{+} X_3$ connection. Here timing of X_2 does not predict timing of X_3 because it is not on the causal pathway. e) Example freeform interventions that are not well spaced. Rows denote variables, hand symbols and thick borders denote interventions. Dashed gray lines show the actual causal influences. f) As in e) but interventions more widely spaced. g) Interventions in a cyclic network. h) Including "blocking" interventions that temporarily prevent activations of the blocked variable and so break feedback loops.

to Gamma($\mu = 2y, \alpha = 2$), doubling the time we expect to wait until seeing X_3 again and introducing a temporary dependence between latest X_2 and the next X_3 (see Subfigure 1.2a, iii). Bramley, Gersten-

berg, Mayrhofer, and Lagnado (2018) provide the mathematical details for estimating model parameters and incorporating base rates and failure rates (similar to causal strengths in CBNs), as well as modelling conjunctive or disjunctive combined causal influences. However, here we simply highlight at a high level what we learn about causal cognition by applying this framework to human judgment and intervention patterns.

1.3.3 Modeling inference

Unlike the CBN case, Bramley, Gerstenberg, Mayrhofer, and Lagnado's (2018) approach captures how timing information informs structure judgments across potentially open ended learning periods, without the necessity of independent trials. If a learner already has a strong domain prior, for instance that a particular causal relationship will take certain length of time to work, this will drive their judgments from time information in a straightforward way, they will only assume an interval is causal if it is close to the expectation. However, different structures map inter-event intervals to different causes-effect delays, meaning that even without a specific prior expectations about delays, the ability of candidate structures to parsimoniously account for the data can be compared. If a learner starts out with no expectations about how long or how reliable causal delays will be — mathematically, if they start out with an uninformative "improper" prior on μ and α — Bayesian Ockham's razor will still favor whatever structure can assign the highest likelihood to the patterns of time dependence observed between variables, with the fewest separate delay parameters (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018). For example, the reliable $X_1 - X_2$ intervals in Figure 1.2b will favor models that include a generative $X_1 \xrightarrow{+} X_2$ connection since this can explain the activations of X_2 better than a model presuming them to be exponentially distributed to caused by any of the other variables.

Bramley, Gerstenberg, Mayrhofer, and Lagnado (2018) show that people spontaneously make structure judgments from event data, broadly in line with the normative framework described above. In three experiments, participants were asked to judge which of a range possible causal structures best describe the activations of components of a set of mystery devices. In one experiment, participants were presented with four video clips of each mystery device, each showing all of the devices' components activate over 3-4 seconds. Each video begins with the activation of the root component (X_1) followed by the activation of three other

output components (X_2 and X_3 and X_4) in varying orders with varying intervals. Participants not only ruled out structures that could not have produced all of the observed patterns, but also assigned more probability to structures to the extent that could reliably have produced the observed delay patterns.

1.3.4 Modeling interventions

Just as with the CBN-packaged observational data in Figure 1.1b, pure observations of events in time cannot unambiguously reveal causal structure. For example, it could be the case that the tight temporal relationship between X_1 and X_2 in Figure 1.2b comes about because the variables share an unmeasured cause that brings about X_1 faster than it brings about X_2. Fortunately, intervention again makes it possible to resolve these kinds of ambiguity. Any temporal dependence between the intervened on variable and activations of other variables suggests the presence of a causal influence.[2] Concretely, for two variables X_i and X_j, if the distribution of inter-event delays $p(t_{\text{do}[X_j],X_i})$ has a best fitting α parameter > 1, this means that X_j hastens X_i relative to its baseline, meaning that X_j (directly or indirectly) influences X_i. A key property of Bramley, Gerstenberg, Mayrhofer, and Lagnado's (2018) modelling framework is that it captures how the sequence of events following an intervention carries information about structure. Post-interventional event *order* places constraints on what could have actually caused what on a given occasion. But beyond this, the causal delays inherit time variability from their parents relative to interventions. Thus, even if an intervention causes the same activations in the same order each time, if the timing of one of these downstream events carries information about another later one, this tells us something about the causal ordering of the variables. Figures 1.2c and d visualize evidence produced by interventions on the root component of a chain and a common cause device. The device in Figure 1.2i is in fact a fork where $X_3 \xleftarrow{+} X_1 \xrightarrow{+} X_2$ with one fast connection $X_1 \xrightarrow{+} X_2$ and one slow one $X_1 \xrightarrow{+} X_3$. Here the $X_1 \xrightarrow{+} X_2$ and the $X_1 \xrightarrow{+} X_2$ delays are uncorrelated. However below we see a $X_1 \xrightarrow{+} X_2 \xrightarrow{+} X_2$ chain structure with two fast connections, one unreliable connection from $X_1 \xrightarrow{+} X_2$ and another reliable connection $X_2 \xrightarrow{+} X_3$. This is revealed by the data for the chain structure, a late X_2 tends to mean a means late X_3 as well.

Prevention is a little more complicated:()

[2] For this to hold interventions must be performed at random or prechosen intervals (thus are unrelated to ongoing activity).

We can think of the propagation of variability through the system as acting akin to a game of "broken telephone" in which the accumulation of errors in a signal (here in the lateness or earlyness of the event timing) can reveal the sequence of message passers. As with broken telephone, if there is no noise (if everyone passes the message perfectly; or if all causal relationships are perfectly reliable) there is noway to tell in what order the message was passed on. But, a moderate about of noise can lead to a "blessing of variability" effect. Bramley, Gerstenberg, Mayrhofer, and Lagnado (2018) show that people are sensitive to this subtle signal, favoring chain structures as for a range of scenarios in which a mediating variable carries information about a distal effect X_3 and fork structure when it does not.

1.3.5 Choosing when and where to intervene

Bramley, Gerstenberg, Mayrhofer, and Lagnado's (2018) approach is applicable to situations where variables may activate or be intervened on multiple times during a single learning period, and in which the underlying structure might contain cycles. Thus, Bramley, Mayrhofer, et al. (2017) also used the framework to explore learners choices of interventions in an extended interaction with a causal device of interest. Instead of giving participants a sequence of independent trials in which to set variables as in traditional interventional learning studies, participants were given 45 seconds to interact with a dynamic causal system and perform up to 6 interventions by clicking on components on the computer screen to activate them.[3] In this setting, learners had to choose both when and where to intervene. Bramley, Mayrhofer, et al. (2017) explored learning about a mixture of acyclic structure (no feedback loops) and cyclic structures (containing at least one feedback loop), and contrasted learning about devices whose causal delays were known to be a reliable 1.5 seconds (standard deviation of 0.1 second) with those that were known to be unreliable (with standard deviation 0.7 seconds). Causal connections had a failure rate of .1 and there were no exogenous activations aside from the learners' own interventions. Consistent with the predictions of the normative model sketched above, Bramley, Mayrhofer, et al. (2017) found that participants were able to identify the majority of the causal connections based on the delayed activation

Maybe by time it comes out we'll have a journal ref

[3] See `https://neilrbramley.com/experiments/it/videos/example_trial.html` for a video of a trial.

information they produced, and were more accurate at identifying the structure of the devices when the delays were reliable.

Bramley, Mayrhofer, et al. (2017) found that successful participants tended to leave large and regular intervals between each intervention, especially when the true causal relationships were unreliable. Successful participants also tended to wait longer since the most recent activation before intervening again, suggesting they waited until they had a low expectation that any other confounding activity would occur during their intervention. For example, Figure 1.2e shows a poorly chosen set of interventions that occur close together, making it ambiguous which activations caused which. By contrast, the well spaced interventions in Figure 1.2f make it clear that X_1 is a common cause of X_2 and X_3. Both X_2 and X_3 reliably follow from the two interventions at X_1, but on the second occasion X_2 and X_3 reverse their order of activation, essentially ruling out the $X_1 \xrightarrow{+} X_2 \xrightarrow{+} X_3$ chain hypothesis that was a plausible until that point. No study has explored preventative causation in the domain of continuous time causal events. However, by contra-position to the above, the ideal time to test for a preventative relationship (as in the $X_2 \xrightarrow{-} X_3$ link in Figure 1.2a) should be when one has a strong expectation the effect is otherwise about to occur.

For the acyclic devices tested in Bramley, Mayrhofer, et al. (2017), participants exhibited a substantial preference for intervening on the root components. Unlike than in most CBN based experimental settings, these positive test interventions *are* informative about the relationships between the downstream variables due to the blessing of variability described above. By intervening at the root of the causal device, learners could then observe how the causality propagated through the system making use of ordering and delay correlations to uncover the causal sequencing (e.g., Figure 1.2c).

Bramley, Mayrhofer, et al.'s (2017) experiment is also the first comparison of human learning about acyclic and cyclic devices in continuous time. Participants' interventions on cyclic devices often led to sustained patterns of activation that would continue until a failed connection would cause the system to acquiesce. This meant that participants experienced substantially more events in total even though they tended, reactively, to perform substantially fewer interventions on the cyclic devices. While this resulted in stronger evidence about structure for the cyclic problems according to the normative learning model, the data was also much more computationally taxing to deal with in real time given the variability in

the true causal delays. The pattern in Figure 1.2g demonstrates this. It is clear from the repeated activations that X_1, X_2 and X_3 are related with a feedback loop. However, establishing the exact set of connections is much harder. There are also many plausible patterns of actual causation that could have given rise to this data — that is, there are many ways one could draw dashed lines as in the figure, to attribute each activation uniquely to a unique previous activation or intervention. To estimate the posterior probability distribution over potential structures it was necessary to sum over large numbers of patterns of possible actual causation (Halpern, 2016). Thus, in is not surprising that human learners with their limited cognitive resources, were substantially worse at identifying the structure of the cyclic devices. Bramley, Mayrhofer, et al. (2017) account for these pattern by positing that human learners build up their causal models incrementally, reacting to each new activation by attributing it to a recent intervention or activation. When there are many causal influences going on simultaneously, this approach becomes less accurate.

1.3.6 Summary and discussion

Bramley, Gerstenberg, Mayrhofer, and Lagnado (2018) developed a normative framework for modeling interventional causal learning and representation in continuous time. Application of this framework to experimental data demonstrated that people use temporal delays between events systematically to infer the causal structure that most parsimoniously explains patterns of observed delays. The framework also showed that people are able to choose sensibly when to intervene, using their expectations about ongoing dynamics to wait until it is unlikely that ongoing processes will confound the interventional evidence. The framework captured why people have a preference toward initiating the root component of causal systems — unlike the CBN-packaged data, this more natural temporally extended data often contains evidential signals about the structure of causality as action propagates though the system.

The analyses also shed light on the intuition that cyclic systems are naturally harder to learn reason about. The computational cost of exact inference in Bramley, Gerstenberg, Mayrhofer, and Lagnado's (2018) framework increased with the number of events that might be related, and cyclic systems tended to produce more activations. The interventions investigated by Bramley, Mayrhofer, et al. (2017) can be thought of as "shocks" to the system, where an additional event is created by

the learner. One way that learners might get clearer structural evidence about dynamic systems in real world contexts is by using preventative interventions to "block" rather that just "shock" the system. Figure 1.2h imagines a case in which a hypothetical learner uses a mixture of punctate "shock" interventions that create activations of variables similar to above, with extended "block" interventions that prevent a variable from activating for an extended period. This latter action is analogous to setting a variable to 0 in the CBN setting explored in 1.2. Blocking breaks the feedback in the system allowing the learner to use interventions to identify one part of the structure at a time as they would in the acyclic cases. Future work could explore how people use a combination of such "shocks" and "blocks" to learn about cyclic systems.

Lagnado and Sloman (2004) proposed that real time interventions act like a strong order cue, with events happening shortly thereafter liable to be associated causally with the action. Bramley, Gerstenberg, Mayrhofer, and Lagnado's (2018) framework captures under what conditions this heuristic will work well, but also formalizes how interventions provide structure evidence regardless of whether this holds. If the base rates of variable activations are low relative to the length of actual causal delays, interventions that activate the cause will tend to create a reliable order cue. That is, effects of the intervention will frequently be the next activations to occur as they are in Figure 1.3e. The fact that participants struggled when there was a dense flurry of activations in Bramley, Mayrhofer, et al. (2017) is consistent with the idea that people really do rely on this kind of cue to build up their hypotheses. However, the normative framework also captures how inference is possible even when this does not hold. Wherever interventions affect the delay distribution for another (e.g. hastening or delaying it relative to its baseline) this is evidence for some form of causation.

Unlike the CBN, a learner who forms a time-extensive representation is also positioned to make real-time predictions about ongoing dynamics (Clark, 2013). This allows them to control for their expectations about ongoing dynamics in choosing when to intervene. This was evident in Bramley, Mayrhofer, et al.'s (2017) experiment in which successful participants would wait for activity to die out before intervening again.

One weakness of Bramley, Gerstenberg, Mayrhofer, and Lagnado's (2018) approach is that normative inference scales rapidly with the number of events experienced, meaning that cyclic or densely connected structures producing large numbers of events become very hard to evaluate. One has to reason about many paths of actual causation playing

out simultaneously, and deal with potentially large numbers of possible paths of actual causation. Fortunately, as the number of events becomes unmanageable for reasoning at the level of event-to-event cause–effect mappings, one can start to reason instead about whether activations and interventions affect the *rates* of occurrence of other variables. Pacer and Griffiths (2015) model this setting, reanalyzing an experiment from Lagnado and Speekenbrink (2010) in which the occurrence of several types of "seismic wave" could temporarily alter the rate at which earthquakes occur. This approach matches with a range of findings suggesting that beyond a small number of unique entities, people switch to approximate counting, known as subitizing, rather than attempting to focus on each entity exactly (Mandler & Shebo, 1982).

Another limitation of this approach is that it is built on the assumption that causality relates to point events. This idealization seems reasonable for processes where the causal influences are occasional and sudden — such as spiking neurons, earthquakes, gunshots, explosions and so on. However, many other causal influences are continuous in time, and many causally relevant variables can take a continua of values with no transition or state being causally privileged. Thus, in the next section we introduce another new framework able to handle learning and inference about variables that evolve and effect one another continuously in time.

1.4 Incorporating time: Continuous variables

While events are often a natural level for reasoning about the world, they need not be construed as discrete or punctate, and change need not thought of as being constituted by *events* at all. Many variables change continuously over time without clear change points, and their consequences can often "linger on and mix with the effects of other actions" (Jordan & Rumelhart, 1992). This means that in many contexts, better causal evidence can be had, and finer grained causal predictions can be made, if one represents the causal world as a continuous system relating continuous variables. Many real world applications of causal inference methods focus on this kind of data. Financial forecasting, climate research, and many aspect of structure estimation in neuroscience involve reasoning about causal influences between continuous variables. The most well established of these approach is, so called "Granger causality", which uses time delayed regressions to assess one variable predicts the later values of another (Granger, 1969). However, this approach suf-

fers from the same limitations of any observational method for assessing causality.[4] Frequently variables *Granger cause* one another without truly causing one another, due to their sharing influences of some unmeasured common cause. For example, because neural activity waxes and wanes due to natural circadian rhythms, regressing one neural signal on another, even with a short delay, will often result in a significant correlation because both measurements share patterns of variation that stem from the brain's overall energy consumption rather than from specifically directed communication. A number of studies have also explored how people learn continuous valued functional relationships more generally, but not in the context of inferring causal structure (Griffiths, Lucas, Williams, & Kalish, 2009; Pacer & Griffiths, 2011; E. Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017). So, in this last section, we survey a new framework for modelling interventional causal learning about continuous valued variables that influence one another in continuous time.

1.4.1 Representing continuous causality

Davis, Bramley, and Rehder (2018) recently developed a new framework for modelling networks of continuous time causal influence. Their approach is based on the Ornstein-Uhlenbeck (OU) process (Uhlenbeck & Ornstein, 1930). An OU process is a stationary Gaussian Markov process capturing random motion that tends to revert to some mean value. Thus, OU processes describe something like Brownian motion that is subject to a corrective force that increases the further the variable strays from its mean. OU processes have been used to model a variety of phenomena including physical (Lacko, 2012) and financial (Barndorff-Nielsen & Shephard, 2001) systems as well as attention during multiple object tracking (Vul, Alvarez, Tenenbaum, & Black, 2009). Davis, Bramley, and Rehder's (2018) insight was to treat all the relationships in a causal network as simultaneously evolving OU processes, such that each variable is noisily attracted to some function of the most recent states of its cause(s).

In a normal OU process, the change in X_i from time t to $t{+}1$ is defined as

$$\Delta X_i^t = N(\theta(\mu - x^t), \sigma) \tag{1.1}$$

[4] Although, see (Eichler & Didelez, 2010) for preliminary work linking interventions with estimates of Granger causality.

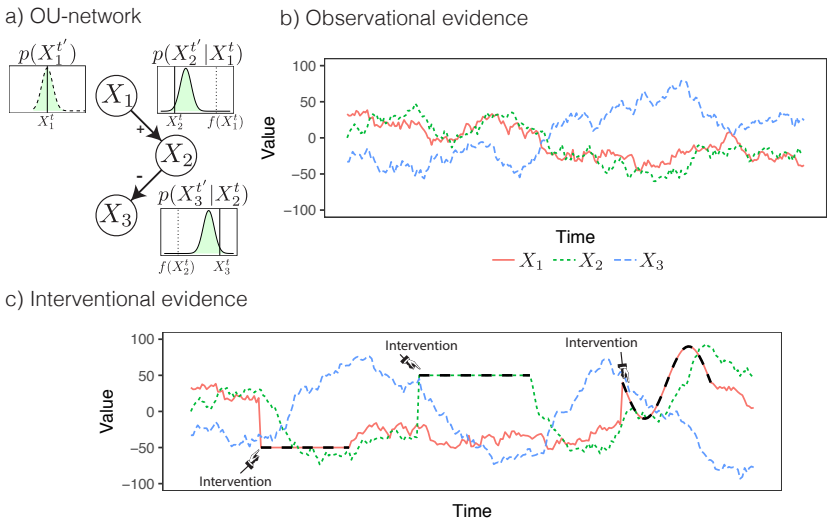

Figure 1.3 Observations and interventions on a causal system relating continuous variables in an OU network. Adapted from Davis, Bramley, and Rehder (2018). Black dashed lines indicate interventions that hold variables in some position or move them systematically.

where μ is the mean that the process reverts to in asymptote, σ is the noise level and θ controls how strongly the process is attracted to its mean. To model ongoing causal relationships, the static μ is replaced by some function of the latest value of the putative cause variable X_j^t. Davis, Bramley, and Rehder (2018) focus on linear relationships, with a single multiplicative parameter w_{ji}, such that the updated value of $X_i^{t'}$ is given by

$$X_i^{t'} = X_i^t + N(\theta(w_{ji} \cdot X_j^t - X_i^t), \sigma). \tag{1.2}$$

Thus, if $w_{ji} > 0$, X_i will be attracted to some positive fraction of the value of the latest value of X_j. In line with the previous sections, we call this a "generative" connection as it implies that increases in X_j cause increases in X_i. If $w_{ji} < 0$, X_i will be attracted to some negative fraction of the latest value of X_j, with X_i tending to decrease as X_j increases and visa versa. Values of w_{ji} between -1 and 1 will make X_i undershoot the absolute value of X_j, while values greater than 1 or less than -1 will make X_i overshoot. There is no causal relationship if $w_{ji} = 0$ or $\theta_i = 0$.

In the former case, the value of X_i simply trends toward zero, and in the latter its movement is a classical random walk.

Davis, Bramley, and Rehder (2018) assume multiple causal influences sum such that

$$\mathbb{E}(\Delta X_i^t) = \theta\Big(\sum_{j=1}^{n} w_{ji} X_j^t - X_i^t\Big) \tag{1.3}$$

and

$$p(\mathbf{w}, \Theta | \Delta X_i^t, X_j^t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\Delta X_i - \mathbb{E}(\Delta X_i^t))^2}{2\sigma^2}} \tag{1.4}$$

where $\mathbf{w}$ collects all w_{ji} parameters for $i, j \in N$ and similarly for Θ and θ_i. Figure 1.2a shows an example continuous variable network $X_1 \xrightarrow{+} X_2 \xrightarrow{-} X_3$ analogous to those explored in Sections 1.2 and 1.3. In it, X_1 has no causes ($\theta_1 = 0$). X_1 is then a generative cause of X_2 ($\theta_2 = .1, w_{21} = 1.25$) which is a preventative cause of X_3 ($\theta_3 = .1, w_{32} = -1.25$).

Figure 1.3b shows example data generated by the network in Figure 1.3a. The root component X_1 (red) drifts around randomly, while X_2 (green) chases noisily behind X_1, and X_3 (blue) chases the inverse of X_2. As with previous settings, it is not possible to tell for certain what the right causal relationships are based on this purely observational data. This is partly because all three variables could be effects of some common cause with different time lags (i.e. with different θ values). Additionally though, the sustained corrective causal influences mean the downstream variables X_2 and X_3 rapidly asymptote to be close to their target values, making it hard to see which moved first unless X_1 happens to move dramatically.

One interesting property about these continuous time networks that is not captured by the event networks in Section 1.3 is their ability to produce rich emergent feedback dynamics of the kinds observed all over the natural world. Figure 1.3c–e give examples of cyclic networks with very different emergent behavior. Figure 1.3c shows an inhibitory feedback loop ($w_{21} = w_{12} = .5$) where both variables trend toward zero no matter where they start. Figure 1.3d shows an excitatory feedback loop ($w_{21} = w_{12} = 2$), where both variables trend toward either positive or negative infinity. Finally, Figure 1.3e shows an oscillatory ($w_{21} = 2, w_{12} = -2$) feedback loop where variables chase one another up and

down in a wave of increasing magnitude. Given the ubiquity of emergent complex behavior in nature, it is counts in the favor of these simple networks that they are able to produce complex emergent dynamics. However it remains to be seen how uniquely observation of such high level dynamics reveals the internal structure of the system as it may be that there are multiple ways of setting up a system to exhibit the same complex dynamics.

Note I've not added subplots for this yet. Its my favorite part of the cogsci paper but perhaps it doesn't belong here?

1.4.2 Modeling inference and interventions

Davis, Bramley, and Rehder (2018) devised a simple experiment to explore human interventional learning about continuous valued variables in continuous time. As in the studies discussed in Sections 1.2 and 1.3, participants were tasked with identifying the structure of a number of causal devices through interacting with them on the computer screen. As with Bramley, Mayrhofer, et al. (2017) they were given 45 seconds to interact with the each device. However, rather than clickable nodes, the variables were represented by vertical sliders. The underlying network could contain generative $w = 1$ and preventative $w = -1$ connections and feedback loops. Unless intervened on, the sliders' positions updated 10 times per second appearing to jitter and track up and down across the (-100)–(+100) range as in Figure 1.2b. Intervening on a slider was achieved by clicking on it, which instantly moved it to the clicked on value, and by hold-clicking participants could keep the variable at a desired location or drag it around smoothly within the sliders' range.[5]

Through interacting with the systems, participants were able to identify the large majority of the connections. As in the other causal learning experiments discussed, the most frequent error was a failure to distinguish direct and indirect causal influences. Participants would often mark direct connections from intervened on variables to indirect effects. By marginalizing over possible $\mathbf{w}$, Θ values under each possible structure, with σ assumed to be known, Bayesian inference could again be used to estimate a posterior for each trial and used as a normative standard against which to compare participants' judgments. Davis, Bramley, and Rehder (2018) compared full Bayesian inference against a local computations model that assumed learners focused on one pair of variables at a time rather than the whole structure at once, finding that around

[5] See `https://neilrbramley.com/experiments/ctcv/demo.html` to try an example trial.

two thirds of participants were better described by this heuristic (cf Fernbach & Sloman, 2009). The normative analysis also revealed that with only one exception, participants interventions dramatically increased the strength of the available evidence.

The role of interventions in this continuous context is intuitively somewhat different from in the previous settings. Because the variables are capable of taking a wide range of values but naturally move only a little per timestep, interventions allow learners to inject dramatic swings and signals into the system, or hold variables at extreme or unusual levels (Figure 1.3, Davis, Bramley, & Rehder, 2018). For example, Figure 1.3c shows a learner performing several dramatic interventions. First they intervene to hold X_1 at 50 for (here, for 40 "timesteps"). This results in a marked increase in X_2 and, with more noise and lag, a reduction in X_3. They then intervening to hold X_2 at -50 which only increases X_3, while X_1 returns to moving randomly. Third, they intervening on X_3 in a sinosoidal pattern leads to a (lagged and noisy) sinosoidal pattern in X_2. While Davis, Bramley, and Rehder (2018) do not analyze participants precise intervention patterns, a large proportion appeared to involve moving and holding variables at extreme values, or rapidly dragging them up and down the full range of the sliders. It is worth noting that in this setting interventions naturally act as both "shocks" and "blocks". Intervening on a variable for an extended period not only takes over its state but also blocks it from participating in feedback dynamics.

1.4.3 Continuous time control

The interventions participants could performed in Davis, Bramley, Gureckis, and Rehder (2018) were often extended in time meaning participants likely to react to the movements of the other variables during their intervention, for example manipulating one variable in such a way as to keep another effect variable in a particular position. This reactivity brings the interventional learning problem increasingly close to an adaptive or "dual control" problem (Feldbaum, 1960; Guez, 2015; Klenske & Hennig, 2016; E. Schulz, Klenske, Bramley, & Speekenbrink, 2017). This means one faces the "dual" problem of learning how the system works while already attempting to control it. For example, learning to play tennis takes place, largely, while attempting to play tennis. At first our shots are wild and do not go where we intend. But, we slowly learn to adapt our swing to different angles and speeds of the incoming ball hopefully building a causal control model of tennis in the process. A key ques-

tion therefore, is to what extent people will spontaneously learn causal structure while attempting to master the control of a causal system of continuously related variables. To investigate this, Davis, Bramley, Gureckis, and Rehder (2018) used their OU driven causal models as a class of control problems. Their task was similar to to that in Davis, Bramley, and Rehder (2018), except that participants interventions were restricted to one "control" variable, and limited single step increments up or down. Participants' goal was to keep another "target" variable in a reward region. To master this task in a model based way, a participant would have to learn the structure of the network and use this knowledge to plan the actions that maximize the time the target variable spends in the reward region. For some simple structures explored by Davis, Bramley, Gureckis, and Rehder (2018), it was enough to exercise model-free control (Dayan & Berridge, 2014). This means simply learning a mapping directly from the control variable to the target variable, without building a model of the relationships between the variables. As it turned out, some structures that involved a mixture of generative and preventative connections were particularly hard to control, requiring the participant not just to understand the causal structure of the system, but also to plan their model based control strategy many frames ahead. In these systems, the only effective way to cause the target variable to enter the reward region involved a periodic oscillation of the control variable. Overall participants learned to control all but these structures well above chance, and learning profiles were broadly consistent with a causal model based controller but also with several other control systems including a deep neural network. Thus future work will need to carefully probe what representation participants learn during a control period.

1.4.4 Summary and discussion

In summary, networks, made up of OU related continuous variables provide a powerful new representation for fine grained continuous time causal structure as well as an interesting test bed of control environments. They map onto real world problems in intuitive ways and display the kinds of emergent properties we see in real world dynamic causal systems. These networks also have a convenient mathematical properties that allow for normative inference and the assessment of the evidential value of interventions. OU processes have the properties of being Gaussian, Markovian, and stationary all of which make likelihood estimation closed form and straightforward (Uhlenbeck & Ornstein, 1930). By hav-

ing all nonstationarity be due to the changing target values determined by the a variable's causes, the resulting system is straightforward to analyze in spite of its rich emergent dynamics. People were able to learn robustly through intervention in this setting, suggesting that they are not limited to reasoning about causality at the level of independent trials or even of real time event cascades.

Interventions in the CTN environment can be used to create dramatic signals that are hard to mistake if they appear in the dynamics of downstream variables. Participants were generally sensitive to this in that they used dramatic swings or held variables at extreme values, which allows for maximally strong causal signal that is easily spotted as it propagates to other variables. However, the richness and immediacy of the evidence also seemed to push participants toward adopting a local strategy focusing on pairs of variables at a time and so struggling to infer the correct structure in the case of indirect causation. As with the event networks, a way to use interventions to get clearer local information in this setting would be to allow combination interventions in which one variable is held in place while another is wiggled around. For instance holding X_2 stationary and moving X_1 up and down in Figure 1.3 would make it easy to establish that X_1does not directly influence X_3 except through X_2. Intuitively, this is a very natural kind of intervention to perform in everyday life, but one that is not captured by the simpler and more abstract notions of intervention and causal influence embodied by the CBN or even the delay networks in Section 1.3. One analogous case though is Steyvers et al. (2003), an early study looking at interventional learning in a CBN setting where aliens read one another's minds. In Steyvers et al.'s (2003) study, aliens' spontaneous thoughts were three letter strings using three different letters of the alphabet ("TUS" or "POR"). However interventions always make them think "ZZZ", which they never normally think. Thus when another alien was revealed to be thinking "ZZZ" it was possible to be very sure that this was effect of the intervention. This highlights a key property of real world interventions that is hidden in the simpler representations. Sometimes it is possible to insert a unique signal with an intervention by setting a variable to a highly unique value or moving it in a highly unusual way. The subsequent inferences is then a kind of message detection, where one looks for traces of the original signal reappearing in other variables.

1.5 General Discussion

There are a number of reasons why CBNs provide a good starting point for thinking about causal cognition. However, we have argued that there are also fundamental reasons why they do not provide a fully adequate account. CBNs represent time only as a partial ordering of influence, while real learning contexts demand a deeper sensitivity to time's arrow. We highlighted recent work introducing two new formalisms using them to explore the ways in which human causal learners are sensitive to time, both in informing their causal beliefs but also in timing their interventions and control. The first representation, developed by Bramley, Gerstenberg, Mayrhofer, and Lagnado (2018), captured causality relating events in time and showed how a preference for reliably timed causal influences can guide structure inference, as well as how sensitivity to the possibility of delayed effects shaped participants intervention behavior. The second representation we discussed was the delay network developed by Davis, Bramley, Gureckis, and Rehder (2018). This latter approach captured causality at an even finer grain, modelling it terms of continuous influences between continuous valued variables. Comparison of human learners in a setting involving these factors again showed that people are capable of interpreting rich continuous evidence as well as using the broad support of the relevant variables to generate distinctive extended and controlled interventions whose signatures could be easily tracked as they propagate through the system.

1.5.1 Interventions in rich domains

While much of the work on intervention selection has taken the perspective of intervention selection as a problem optimal experimental design (Fedorov, 1972). Exploring richer interventions and learning of richer causal representations seems to require a different perspective. Interventions in richer contexts inject signals into causal systems, similar to how a plumber might pour dye down a sink while trying to figure out network of pipes in an old house. Wherever the dye shows up must be downstream of the sink, and the length of time it takes to get there says something about the network of pipes in between. When the propagating signal is not very distinct (such as the time of occurrence of an event as in Section 1.3, or the final states of the variables after a causality has been and gone (as in Section 1.1, collecting multiple approximately independent trials is still important. However, the learner again must be

able to use their knowledge about the domain to create situations that are approximately independent and identical.

One other consideration is that punctate interventions, or "shocks" allow a dynamic system to continue to play out as before, while sustained interventions also act as "blocks", and stop feedback loops. This is valuable in that it can break a cyclic structure learning problem into several acyclic ones. For instance, one can first check whether X_i affects X_j by performing a sustained intervention on X_i before checking whether X_j also affects X_i. However, this approach will shortcircuit the natural dynamics that might themselves be a useful source of evidence. As an analogy, one might contrast the punctate (and probably very distructive) intervention of throwing a coin into a washing machine while it is spinning to a more systematic sustained intervention in which one olds and turns one of the gears in the mechanism and observes what else moves.

1.5.2 The role of theory

All of the representations we considered in this chapter are abstract in the sense that they do not encode anything specific about the mechanisms involved, notably saying nothing about how the parts of the systems are arranged in space. However, we clearly make use of our knowledge of physical mechanisms when reasoning causally (Ahn, Kalish, Medin, & Gelman, 1995; Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018). Indeed, much of the recent work in causal cognition has emphasized the top down role of domain theories on causal inferences (Griffiths, 2005; Griffiths & Tenenbaum, 2009; Lake, Salakhutdinov, & Tenenbaum, 2015). Such domain knowledge is modeled as affecting people's priors about what kinds of structure and parameters are plausible in a given situation. For example basic medical knowledge will tell you that diseases cause symptoms rather than the reverse and that medicines often take on the order of hours to work. Basic knowledge of electronics will tell you to expect causal influences to propagate at the sub-second level rather than across the aeons. Thus, while the current frameworks show how causal beliefs can arise without much specific domain knowledge they are also happily compatible with other factors and knowledge playing into the choice of sensible priors. Another way of thinking of specific domain knowledge is as rich functional forms governing the interactions between variables. For example Bramley, Gerstenberg, Tenenbaum, and Gureckis (2018) explore interventional learning about the physical prop-

erties of objects in simulated physical microworlds. Here learners sophisticated understanding of how objects with different properties normally interact clearly plays into both the judgments they make and the actions they take in service of learning. Interventional behaviors in this domain start to take on recognizable physical form such as shaking and throwing of objects and one another, but still subscribe to the principles exposed here or creating strong interventional signals that propagate through causal relationships (where these are literally magnet-style forces through which the objects in the scenes push and pull one another around). Learners appear to be able to compare the resulting trajectories in time against expectations simulated from their intuitive representation of physics and so back out the relevant properties. In these domains our interventions and our interpretations of them become deeply theory laden, depending profoundly on our beliefs about how the domain works in general (Bramley, Dayan, et al., 2017).

1. Abstraction, discretization & aggregation (i.e. to tabular, CBN-ready data) becomes possible only once we have robust knowledge of when and how to measure effects (i.e. for electrical circuits CBN appropriate because effects are virtually synchronous, other domains we must use our theory to determine the right time windows and sufficient measurements).

1.5.3 When to aggregate

This is all not to say that there is no place for the CBN in causal cognition. The richer representations explored here are useful for navigating causality at an everyday real time scale. But many other phenomena of interest are too long range, and noisy to be identified through close focus on the minutea of change over time, and only become evident as data is aggregated to a much larger scale. Time sensitive representations allow us to make the leap to this higher level of aggregation. Once we have a robust knowledge of when and where to measure effects of medications for example, in our own case we can start gather evidence on a larger scale where time is increasingly abstracted away. For example, reasoning at the level of continuous OU variables might help someone to play tennis, but then when deciding whether to follow a particular strategy it will be more useful to aggregate over multiple games of tennis in which different strategies were pursued, keeping track of how often they work. Essentially we can use our more detailed and time sensitive

theories to determine the relevant time windows and sufficient measurements for studying the subtler and longer scale causal relationships we are interested in establishing. Indeed, it is only through this large scale organized process supported by expertise in the fine grained mechanisms, that humanity has been able to discover the weak relationships between behaviours and medical problems in later life such as the link between smoking and cancer (Gandini et al., 2008), or the lack of one between vaccinations and autism (Verschuur, 1996).

So.. the thought is that intuitive theories/domain knowledge is what ends up allowing us to make the radical simplifications to high level models like CBNs in certain cases. BUT that the raw temporal inference comes first and is more of an important part of human causal learning etc.

In general then the best interventional strategy for a learner depends on what they are interested in learning, how much they already know about the domain, as well as the granularity of available measurement.

1.6 Conclusions

In sum, this recent work is beginning to shed light on the sophistication and nuance involved in everyday causal cognition. We are aware, as was Heraclitus, that one "cannot step into the same river twice"(Barnes, 2013). And so we must rely on our sophisticated model based expectations combined with careful interrogatory actions to learn and exploit the causal world.

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.

Barndorff-Nielsen, O. E., & Shephard, N. (2001). Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 167–241.

Barnes, J. (2013). *Presocratic philosophers*. Routledge.

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.

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.

Bramley, N. R., Gerstenberg, T., & Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 236–242). Austin, TX: Cognitive Science Society.

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

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38. doi: 10.17605/OSF.IO/U9Y4C

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

Bramley, N. R., Mayrhofer, R., Gerstenberg, T., & Lagnado, D. A. (2017). Causal learning from interventions and dynamics in continuous time. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Bruner, J. S., Jolly, A., & Sylva, K. (1976). *Play: Its role in development and evolution.* Penguin.

Buehner, M. J., & Humphreys, G. R. (2009). Causal binding of actions to their effects. *Psychological Science*, *20*(10), 1221–1228.

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.

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.

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

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

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of personality and social psychology*, *58*(4), 545.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204.

Coenen, A., Bramley, N. R., Ruggeri, A., & Gureckis, T. M. (2017). Beliefs about sparsity affect causal experimentation. In *Proceedings of the 39th annual conference of the cognitive science society. austin, tx.*

Coenen, A., Rehder, R., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Cox, D. R., & Isham, V. (1980). *Point processes* (Vol. 12). CRC Press.

Davis, Z., Bramley, N. R., Gureckis, T. M., & Rehder, R. E. (2018). A causal model approach to dynamic control. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Davis, Z., Bramley, N. R., & Rehder, R. E. (2018). Causal learning from continuous variables in continuous time. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 473–492.

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

Deverett, B., & Kemp, C. (2012). Learning deterministic causal networks

from observational data. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Di Blasi, Z., Harkness, E., Ernst, E., Georgiou, A., & Kleijnen, J. (2001). Influence of context effects on health outcomes: a systematic review. *The Lancet*, *357*(9258), 757–762.

Eichler, M., & Didelez, V. (2010). On granger causality and the effect of interventions in time series. *Lifetime data analysis*, *16*(1), 3–32.

Fedorov, V. V. (1972). *Theory of optimal experiments.* Elsevier.

Feldbaum, A. (1960). Dual control theory i–iv. *Avtomatika i Telemekhanika.*

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

Friston, K. J., Bastos, A. M., Oswal, A., van Wijk, B., Richter, C., & Litvak, V. (2014). Granger causality revisited. *NeuroImage*, *101*, 796–808.

Frosch, C. A., McCormack, T., Lagnado, D. A., & Burns, P. (2012). Are Causal Structure and Intervention Judgments Inextricably Linked? A Developmental Study. *Cognitive Science*, *36*, 261–285.

Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A. B., Maisonneuve, P., & Boyle, P. (2008). Tobacco smoking and cancer: A meta-analysis. *International journal of cancer*, *122*(1), 155–164.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1–31.

Gopnik, A., & Schulz, L. E. (2007). *Causal learning: Psychology, philosophy, and computation.* Oxford University Press.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, *71*(5), 1205–1222.

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

Granger, C. W. (2004). Time series analysis, cointegration, and applications. *American Economic Review*, 421–425.

Greville, W. J., & Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, *139*(4), 756–771.

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

Greville, W. J., Cassar, A. A., Johansen, M. K., & Buehner, M. J. (2013). Structural awareness mitigates the effect of delay in human causal learning. *Memory & Cognition*, 1–13.

Grice, G. R. (1948). The relation of secondary reinforcement to delayed reward

in visual discrimination learning. *Journal of Experimental Psychology*, *38*(1), 1.

Griffiths, T. L. (2005). Causes, coincidences, and theories. *Unpublished doctoral dissertation*.

Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with gaussian processes. In *Advances in neural information processing systems* (pp. 553–560).

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. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.

Guez, A. (2015). Sample-based Search Methods for Bayes-Adaptive Planning. *Unpublished PhD thesis*.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, *5*(4), 382–385.

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

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

Holyoak, K. J., & Cheng, P. W. (2011, January). Causal learning and inference as a rational process: the new synthesis. *Annual Review of Psychology*, *62*, 135–63.

James, W. (1890). *The principles of psychology*. Dover (2000 reprint).

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive science*, *16*(3), 307–354.

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.

Klenske, E. D., & Hennig, P. (2016). Dual control for approximate bayesian reinforcement learning. *Journal of Machine Learning Research*, *17*(127), 1–30.

Lacko, V. (2012). Planning of experiments for a nonautonomous ornstein-uhlenbeck process. *Tatra Mountains Mathematical Publications*, *51*(1), 101–113.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 1036–1073.

Lagnado, D. A., & Sloman, S. A. (2002). Learning causal structure. In *Proceedings of the 24^{th} Annual Meeting of the Cognitive Science Society*. Erlbaum.

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *30*, 856–876.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *32*(3), 451–60.

Lagnado, D. A., & Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal*, *3*(2), 184–195.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-

level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lauritzen, S. L., & Richardson, T. S. (2002). Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society*, *64*(3), 321–348.

Malthus, T. R. (1888). *An essay on the principle of population: or, a view of its past and present effects on human happiness*. Reeves & Turner.

Mandler, G., & Shebo, B. J. (1982). Subitizing: an analysis of its component processes. *Journal of Experimental Psychology: General*, *111*(1), 1.

McCormack, T., Bramley, N. R., Frosch, C., Patrick, F., & Lagnado, D. A. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22.

McCormack, T., Frosch, C., Patrick, F., & Lagnado, D. (2015). Temporal and statistical information in causal structure learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *41*(2), 395.

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.

Pacer, M. D., & Griffiths, T. L. (2011). A rational model of causal induction with continuous causes. In *Advances in Neural Information Processing Systems* (pp. 2384–2392).

Pacer, M. D., & Griffiths, T. L. (2015). Upsetting the contingency table: Causal induction over sequences of point events. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality*. New York: Cambridge University Press (2nd edition).

Piaget, J., & Valsiner, J. (1930). *The child's conception of physical causality*. Transaction Pub.

Raiffa, H. (1974). *Applied statistical decision theory*. Wiley.

Rehder, R. E. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *29*(6), 1141.

Rehder, R. E. (2016). Reasoning with causal cycles. *Cognitive Science*, *to appear*.

Rehder, R. E., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264–314.

Rehder, R. E., & Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & cognition*, *45*(2), 245–260.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*(2), 175–221.

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107–1135.

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: observations and interventions. *Cognitive psychology*, *64*(1), 93–125.

Schlottmann, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental Psychology*, *35*, 303–317.

Schulz, E., Klenske, E. D., Bramley, N. R., & Speekenbrink, M. (2017). Strategic exploration in human adaptive control. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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. (2001). Do-calculus?: Adults and preschoolers infer causal structure from patterns of outcomes following interventions. In *Second Biennial Meeting of the Cognitive Development Society*.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. *The Psychology of Learning and Motivation*.

Shannon, C. E. (1951). Prediction and entropy of printed english. *The Bell System Technical Journal*, *30*, 50–64.

Sloman, S. A., & Lagnado, D. A. (2005, January). Do we "do"? *Cognitive Science*, *29*(1), 5–39.

Sloman, S. A., Love, B. C., & Ahn, W.-K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*(2), 189–228.

Sobel, D. M., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory & Cognition*, *34*(2), 411–419.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive Science*, *28*(3), 303–333.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453–489.

Uhlenbeck, G. E., & Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical review*, *36*(5), 823.

Verschuur, G. L. (1996). *Hidden attraction: the history and mystery of magnetism*. Oxford University Press.

Vul, E., Alvarez, G., Tenenbaum, J. B., & Black, M. J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In *Advances in neural information processing systems* (pp. 1955–1963).

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *26*, 53–76.

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learn-*

ing Memory & Cognition, *31*(2), 216–227.

Wolfe, J. B. (1921). The effect of delayed reward upon learning in the white rat. *Journal of Comparative Psychology*, *17*(1), 1.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford University Press.

Young, M. E., & Nguyen, N. (2009). The problem of delayed causation in a video game: Constant, varied, and filled delays. *Learning and Motivation*, *40*(3), 298–312.