Mental Models of Causal Structure in Economics and Psychology

Sandro Ambuehl, Rahul Bhui, Heidi C. Thysen *

March 30, 2026

Abstract

A rapidly growing literature in economics studies how people form beliefs about the causal structures that link economic variables, and what happens when those beliefs are mistaken. We survey this literature and connect it to a large body of related research in cognitive science. After providing an accessible introduction to causal Directed Acyclic Graphs, the dominant modeling approach, we review theory and evidence addressing three nested questions: how individuals reason within known structures, how they estimate their parameters, and how they learn causal structures. We then discuss methodological challenges and describe applications in microeconomics, macroeconomics, political economy, and business.

JEL-codes: D01, D03, D83, D84

*Ambuehl: Department of Economics and UBS Center for Economics in Society, University of Zurich, Blüemlisalpstrasse 10, 8006 Zürich, Switzerland, sandro.ambuehl@econ.uzh.ch. Bhui: MIT Sloan School of Management, 100 Main Street, Cambridge, MA 02142, USA, rbhui@mit.edu. Thysen: Department of Economics, Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway, heidi.thysen@nhh.no.

Contents

1 Introduction

Despite the familiar adage, facts rarely speak for themselves. Economic agents interpret the world through subjective mental models that govern what they learn and how they act. A worker who attributes their low bonus to insufficient effort rather than their manager's hostility will fail to search for a better job. A voter who sees interest rates and prices rising in lockstep blames the central bank for the very inflation it is trying to fight. A firm that advertises heavily every December credits the campaign for holiday sales that might have happened anyway. Thus, even when the facts are not in question, their interpretation may be.

In this paper, we review the burgeoning body of research in economics on subjective mental models of causal structure, and synthesize it with the vast foundational literature from cognitive science.

Traditional research on belief updating in economics typically assumes that individuals know the structure of their environment and asks whether they update beliefs by the correct magnitude. Yet, the structure itself is uncertain in many economically relevant settings. Rather than operating with a perfect understanding, people must rely on simplified mental models, often incomplete or misspecified, to map observations into beliefs and actions. Economists have therefore paid increasing attention to subjective causal models to better bridge the gap between the objective environment and the behavior of its inhabitants. A growing collection of work explores the implications across areas as diverse as political economy, contract theory, behavioral economics, macroeconomics, marketing, and strategic management. This work speaks to phenomena such as cycles of populism, strategic persuasion, systematic cognitive biases, demand for counterproductive macroeconomic policies, misjudgments of product efficacy, and entrepreneurial innovation.

Beginning decades earlier, an extensive parallel literature in cognitive science has investigated how humans learn, represent, and reason about causal structure. One of the enduring puzzles in this tradition is how humans can draw far-reaching inferences from relatively sparse and fragmentary experience—that is, how we appear to 'learn so much from so little'. Rather than learning isolated associations, people often seem to build structured causal representations that constrain interpretation and support generalization. Structure learning, in this view, is a central mechanism by which cognition achieves both efficiency and flexibility.

Despite the shared concerns of economics and cognitive science in this domain, the interaction between the two has been remarkably limited. Cross-field citations are rare, and collaborations rarer still. This disconnect is striking because both fields rely on the same formal framework of *causal Directed Acyclic Graphs* (*DAGs*) developed in computer science and artificial intelligence (Pearl, 2009; Koller and Friedman, 2009). This shared language offers a powerful opportunity for cross-fertilization. For example, cognitive science provides rich empirical evidence on how humans conceive of causal structure, while economics offers sharp theoretical tools for analyzing the behavioral and equilibrium consequences of such conceptions.

This review takes a step toward connecting the fields. We introduce economists to insights from cognitive science, while highlighting open areas that are particularly salient from an economic perspective. For instance, the most prominent economic applications of the framework consider agents who interpret data through misspecified mental models. These applications have remained largely theoretical because the empirical literature in economics on the topic is still nascent, while computational cognitive science has focused more on how people learn the actual structure of the world than on the behavioral consequences of misspecification.

Economists may object that much evidence shows individuals struggle with Bayesian updating even in relatively simple environments (Benjamin, 2019), raising doubts about their ability to reason in settings with unknown causal structure. This concern rests on a particular benchmark for success, namely precise calibration of numerical probability judgments within a well-specified model. But the work reviewed here concerns a much broader set of questions about how agents represent and infer causal structure—which variables they treat as related, in which direction, and through which pathways—and how they use that structure to interpret evidence, predict the effects of interventions, and evaluate competing narratives. The causal perspective helps consolidate a number of stylized facts about belief updating that are not naturally captured by models that treat beliefs as unstructured associations between variables. For example, learning that an outcome has one plausible cause reduces the weight placed on alternative causes (*explaining away*); conditioning on an intermediate variable attenuates inferences about more distant causes (*path blocking*); inferences based on passive observation differ systematically from those based on active intervention (*observation–intervention distinction*); and people readily imagine situations that have never been and never will be observed (*counterfactual thinking*). Indeed, subpar performance in some canonical reasoning problems may reflect the lack of an intuitively coherent causal structure. For instance, when the standard mammography problem is augmented with an explicit causal account of false positives (benign cysts), people are less prone to base-rate neglect (Krynski and Tenenbaum, 2007).

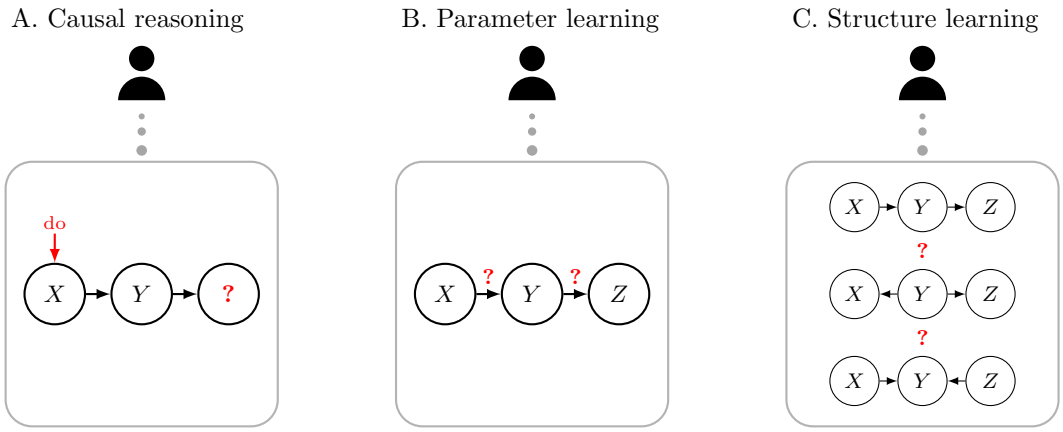
Causal thinking comes naturally enough that even young children display its hallmarks (Gopnik et al., 2001, 2004). Nor is it confined to high-level cognition; Bayesian causal inference unifies many findings about more elementary forms of human information processing, including sensory perception and motor control (Shams and Beierholm, 2022). In fact, Richens and Everitt (2024) prove that any (human or artificial) agent capable of generalizing well across a family of counterfactual data generating processes (DGPs) must have learned, at least implicitly, a causal model of the relevant variables. Hence, causal models are not merely optional refinements to probabilistic reasoning, but are often necessary for approximately optimal decision-making.

The empirical evidence we review is predominantly experimental. Experiments sometimes raise external validity concerns, both regarding stake size and because they typically take place within the very brief time period of a single experimental session. Nevertheless, several studies in experimental economics show that biases in human judgment and reasoning do not disappear (though they are

sometimes slightly attenuated) when stakes are raised by orders of magnitude (Camerer and Hogarth, 1999; Enke et al., 2023). Moreover, research that extends laboratory studies in psychology over timespans of several days or weeks shows that the qualitative results of the single-session studies often generalize (Willett and Rottman, 2021; Zhang and Rottman, 2024; Rottman and Zhang, 2025).

The topics we review extend far beyond the scope of an article-length treatment. We largely limit our coverage of the economic theory literature to learning with misspecified subjective models that take the form of causal DAGs (see also Spiegler, 2020a). Focusing on a different set of questions, Bohren and Hauser (2024) review the literature on more general forms of misspecified learning. Our coverage of formal methods is correspondingly brief. Pearl (2009) provides a comprehensive treatment of causal DAGs from a statistical point of view, to which Pearl and Mackenzie (2018) is an accessible introduction. Koller and Friedman (2009) present the broader formal framework of probabilistic graphical models used to operationalize many of these ideas. The cognitive science literature on causal cognition includes many book-length treatments. To mention but a few, Sloman (2005) offers a gentle introduction, Waldmann (2017) reviews causal cognition more widely, and Griffiths et al. (2024) comprehensively survey Bayesian models of cognition.

Figure 1: The three levels of causal cognition.



The body of this review starts with an informal introduction to causal DAGs (Section 2). Three nested questions then organize the article, as illustrated in Figure 1. First, how do individuals reason within a given causal model, assuming such a model is in place and its parameters are known (*causal reasoning*; Section 3)? Second, how do individuals learn about the magnitude of causal effects within a given structure, and what systematic errors do they make when their beliefs about structure are wrong (*parameter learning*; Section 4)? Third, how do individuals form beliefs about causal structure in the first place (*structure learning*; Section 5)? Each of these sections first presents the relevant theoretical foundations—how idealized learners would draw these inferences—followed by empirical evidence that highlights the regularities, the inconsistencies, and the unanswered questions in each domain. We continue by discussing methodological advances and challenges in measuring beliefs about

structure (Section 6). Section 7 surveys applications of causal cognition across various subfields of economics. The concluding Section 8 highlights several questions beyond the scope of this review.

2 An informal introduction to causal DAGs

The workhorse framework in the literature on causal cognition is the causal DAG. This model encodes possibly stochastic causal relationships among variables of interest. It consists of a set of nodes, each representing a random variable, and directed edges that represent direct causal influence, such as in Panel C of Table 1. An arrow from X to Y indicates that intervening on X may change the distribution of Y, holding all else fixed. Much of the framework's power derives from its ability to represent causal structure in the absence of any functional form assumptions. Fully specifying the distributional assumptions renders the causal DAG a *Causal Bayesian Network*. The framework was originally developed in the 1980s by Judea Pearl in statistics and computer science (see Pearl (2009) for a textbook treatment), introduced to economics by Spiegler (2016), and taken up in a growing psychology literature around the turn of the millennium, including Tenenbaum and Griffiths (2000), Glymour (2001), Gopnik et al. (2004), Sloman (2005), and Lagnado et al. (2007) among others, building on precursors such as Waldmann and Holyoak (1992), Waldmann et al. (1995), and Plach (1999).

Economists unfamiliar with causal DAGs often find it easiest to understand them through the example of systems of regression equations. Consider Panel A of Table 1. It involves standard normal random variables ϵ_i and real-valued parameters β_{ij} for $i, j \in \{X, Y, Z\}$. These quantities define three random variables X, Y, Z. Importantly, we imbue this system with causal meaning by interpreting each equation as how the left-hand-side variable changes if we intervene on the right-hand-side variable. In this way, X is defined as exogenous (since it does not depend on any other variable) while Y and Z are endogenous (since they do depend on other variables).

Panel C of Table 1 shows the DAG representing this system. Each left-hand side variable is a node. To draw the arrows, we consider each variable that appears on the left-hand side of an equation, starting, for instance, with Z. Variable Y appears on the right-hand side of that equation, indicating that it exerts a direct causal influence on Z. To represent this fact in the DAG, we draw an arrow from Y to Z. No other variable appears on the right-hand side of that equation, so we do not draw any further arrows into Z. Next, we consider the equation with Y on the left-hand side. Variable X appears on the right-hand side. Accordingly, we draw an arrow from X into Y. Finally, the equation for X has no variables on the right-hand side. Hence, no arrows point into X. Because the system of equations is recursive, the resulting graph is acyclic. Endowing the DAG with a causal interpretation, which means reading arrows as direct causal mechanisms that determine the effects of intervention, renders it a *causal DAG*.

Table 1: Representations of example probabilistic causal models.

Fully specified probabilistic causal models	
A. Linear Gaussian	***B. Linear binary variables***
Exogenous:	Exogenous:
$X = \beta_X + \varepsilon_X$	$X = L(1, 0; p_X)$
Endogenous:	Endogenous:
$Y = \beta_Y + \beta_{XY} X + \varepsilon_Y$	$Y = L(1, 0; p_Y + p_{XY} X)$
$Z = \beta_Z + \beta_{YZ} Y + \varepsilon_Z,$	$Z = L(1, 0; p_Z + p_{YZ} Y)$

Abstract representations	
C. Causal DAG	***D. Factorization formula***
$X \searrow Y \nearrow Z$	$P(X, Y, Z) = P(Z\|Y)P(Y\|X)P(X)$

Notes: We use $L(a, b; q)$ to denote the lottery that pays a with probability q and b otherwise. The factorization formula represents only the probabilistic information embedded in the causal DAG (conditional independences), but not the causal information.

The literature uses evocative terms to describe properties of nodes in a network: Y is called the *parent* of Z, and Z the *child* of Y. X is a called *root* or *ancestral* node since it has no parents. Z has no children, so Z is a *sink* or *terminal* node. The *ancestors* of Z are X and Y, the variables that cause it directly or indirectly; the *descendants* of X are the Y and Z, the variables that it causes directly or indirectly.

To illustrate the level of abstraction at which causal DAGs operate, Panel B of Table 1 presents a second example with binary random variables, a setting that frequently appears in the economic theory literature (e.g., Eliaz and Spiegler, 2020). Here, X is an exogenous indicator variable, and Y and Z are binary outcomes whose probabilities of success depend on their parents. $L(a, b; q)$ denotes the binary lottery that equals a with probability q and b otherwise. While the success probability of X is constant, that of Y and Z depend on the realizations of other variables, creating dependence between the nodes. We derive the DAG representing this system as before. For instance, because Y appears on the right-hand side of the equation that determines the distribution of Z, we draw an arrow from Y to Z. Despite these differences between the systems in Panels A and B, their DAG representations are identical. The two systems share the same causal structure in the sense of which variables influence which variables, even though they differ substantially in their probabilistic details.

A common question is whether the causal DAG framework is restrictive because it allows causality between variables to flow in one direction only, whereas many models in economics as fundamental as the Keynesian cross involve feedback loops. The standard resolution is to accommodate feedback

loops by unfolding the system over time: rather than representing a cycle between, say, income and consumption, one indexes each variable by period, so that income in period t affects consumption in $t+1$, which in turn affects income in $t+1$ and so on. The resulting graph is acyclic. Such unfolding makes the model dynamic, however. Much of the pertinent literature considers static settings. These can incorporate feedback loops if one replaces the part of the structural system that involves feedback loops with its reduced form, which expresses the endogenous variables as functions of the exogenous variables alone.

Causation vs. correlation Economists are intimately familiar with the conceptual distinction between correlation and causation (sometimes interpreted as prediction and intervention, respectively). The framework of causal DAGs encompasses both.

We represent correlational statements using *conditional probabilities*, written in the form $P(X|Y)$; these capture the *flow of information*. Consider, for instance, the second equation of the causal system in Panel A of Table 1, represented by the link $X \to Y$ in Panel C. Algebraically inverting this equation yields $X = \frac{1}{\beta_{XY}}(Y - \beta_Y - \epsilon_Y)$, which highlights that variation in Y partly reflects variation in X. From our assumption on the causal structure, however, we know that any such information cannot represent a causal effect of Y on X.

To calculate causal effects, we use *interventional probabilities*, written in the form $P(X|\text{do}(Y))$. Such an expression represents the distribution of X if Y is exogenously set to a specific value. To see the difference to conditional statements, consider our example DAG. There, Y does not have a causal effect on X, which we write as $X|do(Y) = X$. Y does, however, have a causal effect on Z. From the structural equation we see that if we set Y to a specific value y, we obtain the distribution of Z after this intervention as $Z|\text{do}(Y = y) = \beta_Z + \beta_{YZ}y + \varepsilon_Z$. This example illustrates the general mechanics of the *do*-operator, which Pearl (2012) calls *graph surgery*: when we intervene on a variable (that is, when we apply the *do*-operator to that variable), we replace the original DAG with a copy that disconnects that variable from its usual causes and replaces it with a constant. The distribution of variables in this altered model correctly captures the post-interventional distribution of all variables. The causal effect on Z from setting Y to 1 rather than 0, for instance, is then given as $P(Z|\text{do}(Y = 1)) - P(Z|\text{do}(Y = 0))$. The formal rules of the *do*-operator are detailed in Pearl (2012). They are consistent with the well-developed intuition about causal interventions that economists typically possess.[1]

Many tools of modern applied microeconomics (such as natural experiments and regression discontinuities) can all be thought of as applications of the *do*-operator in real-world data. In Angrist (1990), for instance, the draft lottery serves as a natural experiment to identify the causal effect of serving in the military. The lottery disconnects the event of serving in the military from its usual

[1] There are some cases where economists' common intuitions might be misleading or specific to a narrow class of models. One example concerns the intuition that total effect = direct effect + indirect effects which is true only in linear models.

causes (such as socioeconomic background) that determine military membership in non-war times and acts as an intervention to determine membership.

Calculations with causal DAGs To perform calculations with causal DAGs, we represent the probabilistic information encoded in the DAG with the *Bayesian network factorization formula*. It relies on the fact that we can write any joint probability distribution as a product of conditional probabilities, in the following way. Consider some n random variables $X_1, ..., X_n$. By the definition of conditional probability, $P(X_1, X_2, ..., X_n) = P(X_n|X_1, ..., X_{n-1})P(X_1, ..., X_{n-1})$. If we apply this definition repeatedly, we obtain the *chain rule of probability*:[2]

$$P(X_1, X_2, \ldots, X_n) = P(X_n|X_1, \ldots, X_{n-1}) \cdot P(X_{n-1}|X_1, \ldots, X_{n-2}) \cdot \ldots \cdot P(X_2|X_1)P(X_1). \tag{1}$$

In our three-equation system of Panel B, we can thus write

$$P(X, Y, Z) = P(Z|X, Y) \cdot P(Y|X) \cdot P(X) \tag{2}$$

The causal structure represented in the DAG allows us to simplify this formula. The crucial information in the system concerns the variables that are *excluded* from the right-hand side in any given equation. For instance, while Z could, in principle, depend on every other variable, the third equation in the system specifies that Z depends only on Y. Therefore, once we know Y, the value of X does not provide any additional information about the distribution of Z. In other words, the third equation guarantees that Z is independent of X conditionally on Y. Formally, $P(Z = z|Y, X) = P(Z = z|Y)$. From the second equation in the system, we know that the distribution of Y depends on X, so that the distribution $P(Y|X)$ cannot be simplified further. Using these insights in equation (2), we obtain[3]

$$P(X, Y, Z) = P(Z|Y) \cdot P(Y|X) \cdot P(X). \tag{3}$$

While we can derive quantitative predictions using the machinery of systems of linear regression equations when all variables are jointly normally distributed, the factorization formula allows us to do this no matter the functional form of the variables' distributions.[4] The statistical literature contains

[2]The symbol P in the foregoing expression represents a probability measure. That is, P is evaluated on sets of values that the collection of random variables $(X_1, X_2, \ldots, X_n)$ can take. When we are interested in specific realizations $x_1, ..., x_n$, we read $P(X_1, X_2, \ldots, X_n)$ as $P(\{X_1 = x_1\}, \{X_2 = x_2\}, \ldots, \{X_n = x_n\})$. Following the canonical convention that uppercase symbols denote random variables and lower case letters specific realizations, the foregoing expression is often written as $P(x_1, x_2, \ldots, x_n)$.

[3]When we apply the chain rule of equation (1), we order the variables. We directly obtain the DAG's factorization into parent-conditional probabilities only if we order the variables such that whenever $X_i \to X_j$ in the DAG, then X_i comes before X_j. If we chose a different order, we would obtain, for instance, $P(X, Y, Z) = P(X) \cdot P(Z|X) \cdot P(Y|X, Z)$, where none of the factors can be simplified using the conditional independencies of the DAG. Moreover, note that the factorization formula is simply a way of describing the joint distribution over all variables, and hence does not retain the causal information embedded in a causal DAG.

[4]Applying it correctly requires appropriately summing over all causal paths by which the intervened-on variable can change the other variables. For example, if we seek to calculate the effect on Z of increasing X from 0 to 1 in the system of Panel B of Table 1, we need to marginalize over Y, that is, to sum over all possible values that Y could take, weighted

additional formula that aid in calculating causal effects in specific cases, such as the *front-door* and *back-door adjustment* formulas to help with confounding variables, and the *mediation formula* to decompose total effects into direct and indirect parts (Pearl, 2009).

3 Causal reasoning

We turn first to causal reasoning: How should idealized learners draw inferences within a given DAG when they know both its structure and the strength of each causal link? And what do humans actually do? We focus on the three archetypal causal structures depicted in Figure 2). They illustrate key phenomena in causal reasoning and have been studied extensively in the empirical literature (Rottman and Hastie, 2014, provide a comprehensive review). The structure $C \rightarrow M \rightarrow E$ is called a *chain*, where C, M, and E are mnemonics for Cause, Mediator, and Effect, respectively. The structure $C_1 \rightarrow E \leftarrow C_2$, in which an effect depends on two causes, is called a *v-collider* or a *common effect* DAG. In the structure $E_1 \leftarrow C \rightarrow E_2$, a single cause influences two effects. It is called a *fork* or *common cause* DAG. We first use these three structures to illustrate three fundamental theoretical concepts: the Causal Markov Condition (how DAGs encode conditional independencies), collider bias (how conditioning on a common effect makes causes dependent), and Markov equivalence (how some DAGs yield identical observational implications). We then outline the empirical evidence concerning causal reasoning.

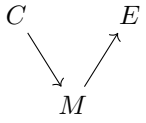
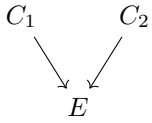
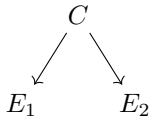
3.1 Theoretical concepts

Chain structure The three-node chain (Panel A of Figure 2) illustrates a central principle of causal reasoning known as the *Causal Markov Condition*. In the chain, the cause C can only indirectly affect the effect E, through the mediator M. Hence, when holding the mediator fixed, changes in C can no longer influence E; the mediator *screens off* the cause from the effect. In correlational terms, the chain structure implies that $\text{cov}(C, E|M) = 0$.[5] The Causal Markov Condition states generally that, conditional on all its direct causes (i.e., parents), a variable is independent of all variables that are not its downstream effects (i.e., descendants). Simply put, once the values of a variable's direct causes are known, information about how those causes themselves came about adds nothing further. A deeper consequence is that, while correlation does not imply causation, it does carry some information about causal structure. For example, if we observe a system in which $\text{cov}(C, E|M) \neq 0$, we can infer

by their respective probability: $P(Z = 1 \mid do(X = x)) = \sum_{y \in \{0,1\}} P(Z = 1 \mid Y = y, X = x) \cdot P(Y = y \mid X = x)$. The factorization formula allows us to evaluate this as $P(Z = 1 \mid Y = 1)\, P(Y = 1 \mid X = x) + P(Z = 1 \mid Y = 0)\, P(Y = 0 \mid X = x) = (p_Z + p_{YZ})(p_Y + p_{XY}x) + p_Z(1 - p_Y - p_{XY}x) = p_Z + p_{YZ}(p_Y + p_{XY}x)$. The causal effect of X on Z is the difference of this expression evaluated at $x = 1$ and $x = 0$, which equals $p_{XY} \cdot p_{YZ}$, which is the product of the effects along the chain.

[5] Formally, the structure implies conditional independence $C \perp\!\!\!\perp E|M$, which is stronger than correlation and is defined even if the variables are multinomial rather than real-valued. Because many economists are more familiar with thinking about about covariance than about independence, we use correlational statements when doing so creates no confusion.

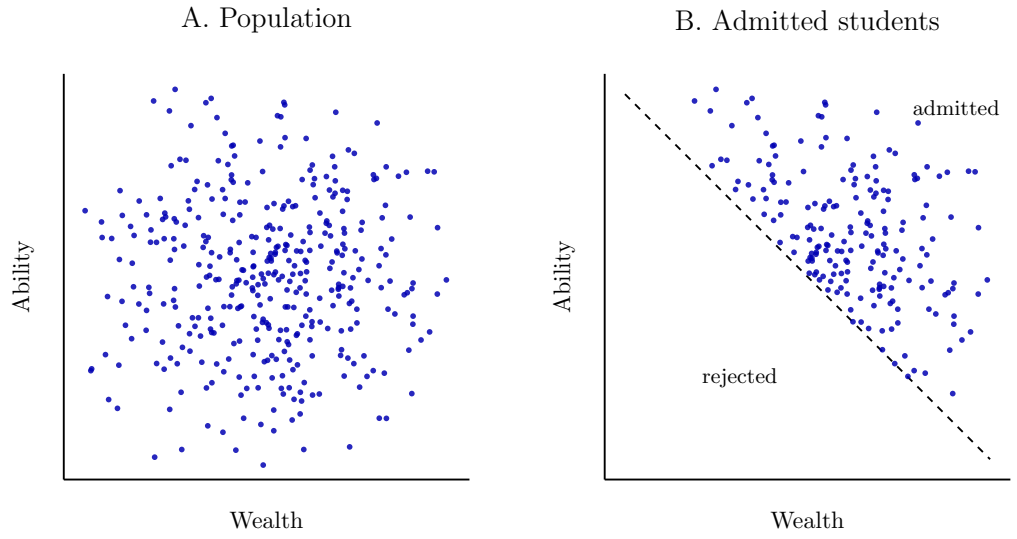
Figure 2: Archetypal causal structures

	(A)	(B)	(C)
Common labels	Chain, Line	Common Effect, v-collider	Common Cause, Fork
			
Example	Salt consumption increases blood pressure which decreases life expectancy	Wealth and ability both determine admission to a selective university	Holiday seasonality drives both high sales and more advertising

that the DGP cannot be a chain. More broadly, any DAG implies a list of conditional independence relationships, each stating that one set of variables is independent of another, conditional on a third (possibly empty) set. Comparing these implications to the conditional independence relationships present in the data allows the analyst to narrow down the candidate set of DAGs that may have generated the data.

Common effect structure Some of the most distinctive implications of causal inference in DAGs arise in common effect structures (Panel B of Figure 2), where multiple causes influence a single effect (such an effect is called *collider node*). In these settings, conditioning can have the opposite effect from the screening-off property discussed above. Rather than blocking associations, conditioning on a common effect *induces* correlations between its causes—a phenomenon known as *collider bias*. Formally, a common effect structure implies that, generically, $\text{cov}(C_1, C_2) \neq \text{cov}(C_1, C_2|E)$ (if C_1 and C_2 are real-valued). While this result may be surprising, it matches people's intuitive predictions. Suppose for example that academic ability and parental wealth are uncorrelated in the population, so that information about one is generally uninformative about the other. Consider now the student body at a university that admits students either on ability or through parental donations. Once we learn that a student has wealthy parents, we become more pessimistic about that student's ability compared to other students at the university, since parental wealth may have compensated for lower ability in the admissions decision. As Figure 3 illustrates, conditioning on the effect (E, having been admitted to the university) creates a correlation between the two otherwise unrelated variables, ability (C_1) and parental wealth (C_2). If, as in the admissions example, collider bias arises from selection, it is known as *Berkson's paradox*. Collider bias also explains other apparent paradoxes, such

Figure 3: Collider bias / Berkson's paradox



Notes: Even if ability and wealth are uncorrelated in the population, they become correlated once we condition on the event of being admitted to a selective university that selects students based on both ability and wealth.

as the celebrated Monty Hall Problem, in which the door Monty opens depends both on the door the contestant selected and the location of the car.[6]

The admissions example illustrates the phenomenon of *explaining away* (sometimes referred to as *causal discounting*), a special case of collider bias which has been studied extensively in the psychology literature (Khemlani and Oppenheimer, 2011). In common effect structures with binary variables and independent or substitutable causes (e.g., where either cause alone can generate the effect), observing that one cause was present reduces the posterior probability that another cause was also present. Intuitively, when one sufficient explanation for an outcome is identified, alternative explanations become less likely. Formally, explaining away occurs when $P(C_1 = 1|E = 1) > P(C_1 = 1|E = 1, C_2 = 1)$. This inequality arises when one cause matters less for producing the effect once the other cause is present (in the sense that the likelihood ratio $P(E = 1|C_1 = 1, C_2)/P(E = 1|C_1 = 0, C_2)$ is decreasing in C_2).[7] Substitutability of the causes is critical; when causes are complementary (for instance, when

[6] In the Monty Hall problem a contestant in a game show chooses one of three doors, behind one of which is a prize (typically a car) and behind the other two are goats. After the contestant's choice, the host, Monty Hall, who knows what is behind each door, opens one of the remaining doors to reveal a goat, then offers the contestant the chance to switch to the other unopened door. The counterintuitive result is that switching is the optimal strategy and doubles the probability of winning from 1/3 to 2/3. As Pearl and Mackenzie (2018) explain, conditioning on the door Monty opened creates a (non-causal) correlation between the two determinants of his choice; it is this correlation that renders switching the optimal strategy.

[7] To see this, note that by Bayes' rule and independence of the two causes, $P(C_1 = 1 \mid C_2 = c_2, E = 1) = \frac{p_{1c_2}\,\pi_1}{p_{1c_2}\,\pi_1 + p_{0c_2}(1-\pi_1)}$, where $p_{c_1c_2} = P(E = 1 \mid C_1 = c_1, C_2 = c_2)$ and $\pi_1 = P(C_1 = 1)$. This expression is increasing in the likelihood ratio p_{1c_2}/p_{0c_2}, so the decreasing likelihood ratio condition $p_{11}/p_{01} < p_{10}/p_{00}$ directly implies $P(C_1 =$

both are necessary to cause the effect) learning that one cause was present can make it more rather than less likely that the other cause was also present.

Collider bias occurs regardless of whether the multiple causes of an effect are independent (as in Panel B) or whether they can causally affect each other (as would be the case, for instance, if the DAG featured an additional edge $C_1 \rightarrow C_2$). Yet, common effect structures in which the causes of an effect are *not* directly causally linked—a class called called *v-colliders* (where the letter v is a graphical mnemonic)—have fundamentally different implications when fitting models to data than colliders whose parent nodes are linked (Section 4). DAGs in which no collider is a v-collider are called *perfect* ('all parents are married').

Common cause structures In common cause structures (Panel C of Figure 2), a single cause leads to two separate effects. The Causal Markov Condition therefore also applies in these networks: holding fixed the common cause, the realization of one effect provides no information about the other effect. Formally, E_1 and E_2 are independent conditional on C.

Importantly, the correlational implications of the common cause structure coincide with those of the chain $E_1 \rightarrow C \rightarrow E_2$, for which the Causal Markov Condition also implies that E_1 and E_2 are independent conditional on C. This is the only independence relationship in either structure.

This example illustrates the core concept of *Markov equivalence*. While distinct DAGs always differ in their full set of *interventional* implications, they sometimes make identical *conditional predictions*—that is, they sometimes entail the same set of statistical associations in observational data. Such DAGs are called Markov equivalent. An immediate implication for structure learning (covered in Section 5) is that correlational information alone cannot distinguish between Markov-equivalent DAGs. Relatedly, when making conditional predictions based on passive observation it does not matter whether one has fit a misspecified DAG to the data as long as that DAG is Markov equivalent to DAG that characterizes the DGP.

Determining Markov equivalence may appear hard because deriving the full set of correlational implications for a DAG is generally laborious. Fortunately, Verma and Pearl (1990) provide a straightforward criterion: two DAGs are Markov equivalent if and only if they have the same *skeleton* and the same set of v-colliders (called *v-structure*). One obtains the skeleton of a DAG by dropping all arrowheads, that is, by making the directed graph into an undirected one. Using this criterion, we can quickly see that the chain (panel A of Figure 2) and the common cause structure (panel C) are Markov equivalent: both have an empty v-structure, and once we drop the arrowheads, they look the same. We can also see that they are not Markov equivalent to the common effect structure (Panel B) which features a non-empty v-structure. Hence, an analyst considering these three candidate structures can, based on observational data alone, infer whether or not the DGP is a common effect structure, but cannot distinguish whether it is a chain or a common cause structure.

$1 \mid C_2 = 1, E = 1) < P(C_1 = 1 \mid C_2 = 0, E = 1)$, which in turn implies $P(C_1 = 1 \mid C_2 = 1, E = 1) < P(C_1 = 1 \mid E = 1)$ by the law of total probability.

Common cause structures play an important role in cognitive science because they give rise to the phenomenon of *cue combination*, a well-studied topic in domains as fundamental as sensory perception (Trommershäuser et al., 2011). It concerns how multiple signals are integrated, based on their reliability, to infer an underlying latent cause, which entails computing $P(C|E_1, E_2)$. This logic is familiar to economists. Multiple noisy observations (e.g., financial indicators, expert assessments, or survey responses) can each be informative of an underlying state, reflecting the aggregation of conditionally independent evidence about a shared cause.

3.2 Empirical evidence

Causal reasoning has been studied extensively in cognitive science, typically using externally provided causal structures wrapped in cover stories. Some studies take inspiration from everyday physical, social, or biological systems, where participants' judgments may naturally recruit existing domain knowledge. When seeking to limit the influence of priors, studies deliberately feature unfamiliar or fictional settings, such as interactions between novel chemicals or science-fiction vignettes about aliens. Studies also vary by whether they verbally describe the probabilities that parametrize the DAGs, present data about individual realizations in tabular form, or have subjects experience a sequence of realizations (called a trial-by-trial design).

Overall, experimental research finds that subjects' behavior is broadly consistent with normative predictions, in the sense that 'when the normative calculations imply that an inference should increase, judgments usually go up; when calculations imply a decrease, judgments usually go down' (Rottman and Hastie, 2014). One important caveat, however, is that subjects' judgments tend to vary less with information and experimental manipulations than predicted. This attenuation is reminiscent of a general tendency for conservative belief updating, possibly for the same reasons that underlie conservative updating in balls-and-urns tasks (Benjamin, 2019; Enke et al., 2024). Studies in which subjects learn probabilities by experience tend to find greater adherence to the normative benchmark than those that describe probabilities verbally. Beyond these general patterns, the literature also documents various deviations from predictions in specific causal structures.

Causal Markov Condition Do experimental subjects' judgments adhere to the Causal Markov Condition? In many cases, people clearly recognize its qualitative implications and reason accordingly. In Ambuehl and Thysen (2024), for instance, subjects could increase their payoff by selecting which of two causal models was consistent with correlational data generated by an underlying DGP, based on verbal descriptions using terms such as 'symptom', 'cause', and 'mediator' as well as graphical depictions. Around two thirds of subjects consistently identified the correct model when doing so required intuiting that fixing the mediator in a chain DAG blocks the influence of the cause on the effect. Relatedly, Sussman and Oppenheimer (2011) show that when people predict an unknown variable from two observed variables in a deterministic setting, the weight they place on each predictor

tracks the conditional independence structure implied by the Causal Markov Condition. In chain and common cause structures, their subjects placed nearly all weight on the screening-off variable and effectively ignored the other variable, even though they weighted them roughly equally when both predictors were independent causes of the target.

These studies consider settings that are either deterministic (Sussman and Oppenheimer, 2011) or feature large amounts of data that eliminate uncertainty about the magnitudes of correlations (Ambuehl and Thysen, 2024). They also present subjects with abstract variables that do not suggest any real-world analogs. However, when the underlying causal networks feature more stochasticity, when subjects are presented with descriptions of causal relations between real-world variables, and when they are asked to internalize these structures rather than choose between them, humans reliably violate the Causal Markov Condition, as Rottman and Hastie (2014) conclude. Rehder (2014), for instance, asked subjects to assume that three variables were linked in a chain DAG. Even when he described the value of the middle node as fixed, subjects judged the effect as being more likely to occur when the upstream cause (the ancestral node in the chain) was present rather than absent, in violation of the screening off property. Similarly, with a common cause structure, subjects used one observed effect to draw inferences about the other effect even when the cause was observed. In a common effect structure with independent causes, subjects tended to treat the causes as correlated despite no apparent reason for this. These patterns do not appear to be driven by time pressure, suggesting that they reflect systematic features of causal reasoning (Rehder, 2014; Kolvoort et al., 2024), and they persisted when causal relations were described between abstract variables rather than real-world entities (Rehder, 2014). In line with a more general tendency (Lejarraga and Hertwig, 2021), some studies in which subjects learn DAG parameters by experience rather than from description find greater adherence to the normative benchmark (Rehder and Waldmann, 2017a), though evidence is mixed (Rottman and Hastie, 2016); see Section 4.2.

Several explanations for Markov violations have been proposed (Rottman and Hastie, 2016). One possibility is that subjects posit an unobserved causal mechanism linking variables that are assumed to be independent in the experimenter's model, effectively enriching the causal structure beyond what is explicitly described (Rehder and Burnett, 2005; Park and Sloman, 2013; Marchant et al., 2023). Consider, for instance, the example $fitness \leftarrow jogging \rightarrow weight\ loss$. Taking this DAG at face value, conditioning on *jogging* eliminates the correlation between $fitness$ and *weight loss*. In the real world, however, jogging is not the only physical activity that can influence these variables; biking, say, has a similar effect. Subjects asked to judge the correlation between $fitness$ and *weight loss* holding constant *jogging* may therefore draw on this richer real-world model, in which variables such as *biking* generate a residual correlation. Such subjects will appear to violate screening off, even though their judgments are perfectly consistent with a more complete causal structure.[8] Relatedly, subjects

[8] In one experiment of Park and Sloman (2013) for instance subjects reason about two consequences of living in Africa (C). In one treatment the two effects are both mediated by the same mechanism: subjects are told that the intense sunlight causes people to wear sunglasses (E_1) as well as long sleeves (E_2). In the other treatment, long sleeves are

may believe that the observed variables exhibit measurement error and are imperfect indicators of underlying states, in which case the conditional independencies implied by the stated DAG may not apply in their subjective model (Rehder and Burnett, 2005). A different account emphasizes associative biases whereby reasoners treat variables as more likely to co-occur when they share causal links, even when such associations should be screened off given the structure (Rehder and Waldmann, 2017b). Yet another explanation focuses on the cognitive process, proposing that people compute approximate inferences by mentally sampling states of a causal network in a biased way (Davis and Rehder, 2020; Kolvoort et al., 2023). Nevertheless, we are not aware of any account that is broadly considered able to accommodate all empirical patterns reported in the literature.

Explaining away Building on the seminal empirical study of Jones and Harris (1967), Morris and Larrick (1995) asked subjects to read an essay supporting Fidel Castro and informed them that there were two possible reasons why it might have been written rather than an essay criticizing him. With a 50% chance, the author was free to choose which view to express; otherwise, the author was instructed to write either a pro-Castro or an anti-Castro essay, with the two assignments being equally likely. Once subjects were informed that the author was in the random-assignment condition, their belief that the author personally favored Castro dropped significantly. The assignment condition thus explained away the hypothesis that the author held pro-Castro views.

With deterministic causal structures, even young children exhibit explaining away (Gopnik et al., 2004). In the classic 'blicket detector' task, children observe that certain objects ('blickets') activate a machine when placed on it, while others do not. Multiple objects are initially presented together, leaving open which of them is causally responsible. When children later learn that one object alone is sufficient to activate the machine, they correspondingly reduce their belief that a second object is also a cause.

In the literature that has followed (see Rottman and Hastie, 2014; Khemlani and Oppenheimer, 2011), explaining away occurs often but not always, and when it does, it is typically attenuated relative to the Bayesian benchmark. Judgments about the probability that one cause occurred generally do move in the predicted direction when information about an alternative cause is revealed, but less strongly than normative models predict. The strength of explaining away varies substantially across individuals and experimental contexts (Sussman and Oppenheimer, 2011), and is sensitive to task features. Like failures of screening off, failures of explaining away seem more common when probabilities are learned from description than from experience (Rehder and Waldmann, 2017a). Failures of explaining away do not, however, seem to be related to time pressure (Kolvoort et al., 2024). Various theories have been proposed to describe these empirical patterns (Khemlani and

framed as malaria protection, and hence are mediated through a different mechanism. Across multiple such examples, they find that in the same-mechanism condition, $P(E_1 \mid C)$ drops from 75.8 to 66.5 when subjects additionally learn that E_2 is absent, in violation of screening-off. In the different-mechanism condition, by contrast, the corresponding drop is a much smaller; from 75.1 to 72.0.

Oppenheimer, 2011), including those which address Markov violations (Rottman and Hastie, 2016), such as *causal elaboration* (in which reasoners consider or invent causes outside the explicit problem description Rehder and Burnett, 2005; Park and Sloman, 2013; Marchant et al., 2023) and biased sampling (Davis and Rehder, 2020; Kolvoort et al., 2023).

Markov equivalence Markov-equivalent DAGs produce identical correlational predictions but differ in their implications for intervention. For such structures, causal direction should therefore be irrelevant when making predictions from observed associations, but should matter when choosing actions. Do people adhere to these principles?

A body of evidence answers this question broadly in the affirmative (Sloman and Hagmayer, 2006; Hagmayer and Fernbach, 2017). Experiment 4 of Hagmayer and Sloman (2009), for instance, used two Markov-equivalent structures, a chain and a common cause over the same three variables. The action served as root cause in the chain and as one of the two effects in the common cause structure; participants were queried about the presence of an auxiliary variable that functioned as mediator in the former case and as common cause in the latter. When the action was merely observed, participants' inferences about the auxiliary variable were identical across structures, consistent with the principle that Markov-equivalent DAGs are indistinguishable from observational data. However, when the action was deliberately chosen or externally forced, they considered the auxiliary variable more likely to be present in the chain condition, as prescribed by the *do*-operator.[9]

However, research on the *predictive/diagnostic reasoning asymmetry* has documented a systematic deviation. By Markov equivalence, the two DAGs $X \rightarrow Y$ and $X \leftarrow Y$ have the same correlational properties. Yet people tend to be more confident in their predictive inferences from cause to effect than in their diagnostic inferences about causes from effects. Tversky and Kahneman (1980) first documented this effect by asking subjects whether it is more likely that a blue-eyed mother's daughter also has blue eyes, or that a blue eyed girl's mother also has blue eyes. Although both conditional probabilities should be equal, the overwhelming majority believed the former to be more likely. Similarly, they believed that predicting a son's height from his father's height would be more accurate than predicting a father's height from his son's height. People thus behave as if they perceive greater predictive strength along the direction of a causal arrow (*predictive reasoning*) than in the reverse direction (*diagnostic reasoning*). In some cases, the effect appears to reflect a tendency to consider alternative causes (e.g., the father's eye color in the blue-eyes example) when reasoning from effect to cause, but to neglect them when reasoning from cause to effect. This difference in the consideration of alternatives is sometimes taken to define the asymmetry (Fernbach et al., 2011, 2010).

[9] Closely related experiments in Hagmayer and Sloman (2009) and Hagmayer and Meder (2013) find similar results in settings where the underlying structures are not strictly Markov equivalent (as they do not contain exactly the same variables) but are observationally equivalent in a broader sense (because some variables are unobserved).

4 Parameter learning

In the previous section, we considered agents who reason within a fully parametrized DAG. What if an agent only knows the causal structure, represented by the DAG, but not the magnitudes or strengths of its causal links? These magnitudes can be inferred based on the available data, paralleling standard statistical approaches. We outline the theoretical concepts before reviewing the relevant empirical evidence (Section 4.2).

4.1 Theoretical concepts

We first discuss parameter learning when the DAG is correctly specified (Section 4.1.1). We then turn to learning with misspecified models (Section 4.1.2), which is more closely tied to the kinds of non-standard behaviors studied in behavioral economics.

4.1.1 Parameter learning with correctly specified models

Fitting a DAG to data If a reasoner has in mind a DAG that describes the causal relations between variables, how can she use data to learn about the strength of these relations? When the variables in a DAG are discrete, doing so is conceptually straightforward. The DAG factorization formula lets us express the joint distribution of all variables as a product of conditional probabilities. Replacing these conditional probabilities with their sample counterparts (the observed proportions in the data) yields the maximum likelihood estimates of the parameters of the corresponding Bayesian network, and thus of the joint distribution it implies (Koller and Friedman, 2009, Chapter 17). This process is called 'fitting the DAG to the data,' 'parameterizing the DAG,' or 'projecting the data through the DAG.'

To illustrate, suppose we wish to fit the collider DAG $X \rightarrow Y \leftarrow Z$ to the data in Table 2. Its factorization formula is $P(X, Y, Z) = P(Y|X, Z)P(X)P(Z)$. We can estimate $P(X)$ from the dataset by counting all observations for which $X = 1$, yielding $P(X = 1) = 0.50$. In a similar way, we obtain $P(Z = 1) = 0.50$. We read off the collider node Y's conditional distribution by focusing on all observations with a given combination of X and Z (that is, within each rectangle in Table 2) and calculating the fraction of those observations for which $Y = 1$. For example, the top rectangle contains the four observations with $X = 0$ and $Z = 0$, one of which has $Y = 1$, yielding $P(Y = 1|X = 0, Z = 0) = 0.25$. Repeating this calculation for each of the four rectangles yields the full conditional probability table for Y. In the present example, this reveals that $P(Y = 1|X, Z) = 0.75$ if $X = 1$ or $Z = 1$, and 0.25 when both $X = 0$ and $Z = 0$. This result suggests that the function that determines the state of Y depending on X and Z (called the *link function* or *connector*) is similar

Table 2: Dataset for DAG fitting

Observation ID	X	Y	Z	
1	0	0	0	X=0, Z=0
2	0	0	0	
3	0	0	0	
4	0	1	0	
5	0	0	1	X=0, Z=1
6	0	1	1	
7	0	1	1	
8	0	1	1	
9	1	0	0	X=1, Z=0
10	1	1	0	
11	1	1	0	
12	1	1	0	
13	1	0	1	X=1, Z=1
14	1	1	1	
15	1	1	1	
16	1	1	1	

Notes: Hypothetical dataset with $N = 16$ observations of three binary variables X, Y, Z. Each rectangle groups the four observations sharing a given combination of X and Z. The highlighted column marks Y, the dependent variable in the collider DAG $X \to Y \leftarrow Z$. Reading off the fraction of $Y = 1$ within each rectangle yields $P(Y|X, Z)$.

to a logical OR function: Y is 'on' with high probability if either input node is 'on,' and with low probability otherwise.[10]

Observe that this inference requires no functional form assumptions; the OR-like relationship is revealed directly by the data. Had the data instead shown that $P(Y = 1|X, Z)$ is high only when both X and Z are 1 but low otherwise, we would instead have inferred an AND-like relationship.

While the dataset in Table 2 was indeed generated from the collider $X \to Y \leftarrow Z$, we can also fit the chain DAG $X \to Y \to Z$ to this data. This DAG is misspecified given the DGP, but the exercise illustrates the mechanics of parameter fitting for different causal structures. The factorization formula for the chain DAG is $P(X, Y, Z) = P(Z|Y)P(Y|X)P(X)$. As before, $P(X = 1) = 0.50$. Because Z is not a parent of Y in this DAG, we estimate $P(Y = 1|X)$ by *marginalizing* over Z, meaning that we sum over all observations with a given value X regardless of their value of Z. We find that $P(Y = 1|X = 1) = 0.75$ and $P(Y = 1|X = 0) = 0.50$. Similarly, we obtain $P(Z = 1|Y = 1) = 0.60$ and $P(Z = 1|Y = 0) = 0.33$. As will become relevant later, we see that interpreting the data in Table 2 through the lens of the chain suggests that raising X increases the probability that $Y = 1$, because $P(Y = 1|X = 1) > P(Y = 1|X = 0)$, which in turn increases the probability that $Z = 1$, because $P(Z = 1|Y = 1) > P(Z = 1|Y = 0)$. Hence, viewing this dataset through the lens of a chain DAG (wrongly, in this case) suggests that X can exert an indirect causal effect on Z.

[10] A closely related formal concept, the *noisy-OR* specification, models each cause as having an independent probability of producing the effect. By implication, if both causes are absent, the effect is absent with certainty. The *leaky noisy-OR* specification generalizes the noisy-OR by allowing the effect to occur with some probability even if no cause is present.

This estimation approach can require a large amount of data. If one has additional information, such as knowledge of functional form, more statistically efficient approaches may be used. An analyst facing continuous variables that are jointly Gaussian, for instance, can estimate parameters with linear regression.[11]

Bayesian learning The above section describes maximum likelihood estimation of DAG parameters. The approach can be naturally extended to a Bayesian framework. Doing so yields not only point estimates but entire posterior distributions over parameters. This information is needed for Bayesian structure learning (Section 5), which asks how probable different causal structures are in light of the data. Implementing this approach requires specifying a prior. A highly tractable, and therefore common, choice is the Dirichlet-Multinomial specification, which assumes that each variable has finitely many possible realizations and places a Dirichlet prior over the conditional probability parameters (see Section 17 of Koller and Friedman, 2009). Practically, inference then boils down to supplementing the observed data with hypothetical prior data (called *pseudocounts*), and fitting the DAG to this combined dataset.

4.1.2 Parameter learning with misspecified models

As we have just seen, parameter learning is conditional on a causal structure. If that structure is misspecified, agents will be misled by the data even when learning is otherwise well-behaved. Therein lies a major appeal of this approach for behavioral economics. As Spiegler (2016) writes, the framework of fitting misspecified causal DAGs to data '*provides a "general recipe" for transforming a standard rational expectations model into an equilibrium model with nonrational expectations.*'

Because many distinct DAGs can be fit to the same DGP, an immediate question is which misspecifications matter. Some subjective DAGs may serve as good enough 'as-if' models, while others generate substantial errors. The answer depends on whether the decision maker (henceforth *DM*) is interested in using observations to form non-interventional beliefs, or whether she instead seeks to choose an intervention. A DAG's correlational implications suffice for the former case but not for the latter. We will discuss these two cases in turn.

Non-interventional beliefs with misspecified DAGs For the purpose of conditional predictions—that is, forming expectations conditional on observed variables—a causal DAG is completely characterized by the list of conditional independence relationships it implies (see Section 2). DAGs that represent the same list of conditional independence relationships are called *Markov equivalent*. Therefore, as long as the agent's misspecified DAG is Markov equivalent to the DGP, her predictions will be just as good as they would be if she had fit the correct DAG.

[11] The joint Gaussian assumption implies that there are no interaction terms in the conditional mean, because if X is an n-dimensional vector of jointly Gaussian random variables, then $E(X_i|X_J)$ is linear in X_J for all $i \in \{1,\dots,n\}$ and $J \subseteq \{1,\dots,n\}$.

If the agent fits a DAG that is not Markov equivalent to the DGP, she will make mistaken predictions for some nodes and some parameterizations of the DGP. However, these errors may be concentrated on variables of no direct interest to the agent. Is there a way to determine whether the agent's predictions will be wrong in a 'relevant' way? In fact, there are two such results that differ by the definition of 'relevant'.

First, Spiegler (2020b) considers a setting with n nodes, $X_1, \ldots, X_n$. He assumes that the DM cares about the subset of variables $S \subseteq \{X_1, \ldots, X_n\}$. His Proposition 2 shows that the agent will have correct beliefs about the joint distribution of all variables in S if and only if S is an ancestral clique in some DAG G that is Markov equivalent to the DM's subjective DAG. A set of nodes S is an *ancestral clique* in a DAG G if G directly connects all variables in S to each other, and any variable that G claims causes any other variable in S is itself in S. To illustrate, in none of the DAGs in Figure 2 do all three nodes form an ancestral clique, as each DAG involves two variables that are not directly connected. The sets $\{C, M\}$ in the chain (Panel A) and $\{C, E_1\}$ in the fork (Panel C) constitute ancestral cliques, but $\{C_1, E\}$ in the collider (Panel B) does not, since it includes E while omitting one of E's causes, C_2.

Second, as a corollary of the previous condition, Spiegler (2020b) considers a DM who is not interested in correlations between variables but merely cares about each variable's marginal distribution—for instance, if she is only interested in the mean of each variable. With this narrower definition of relevance, Spiegler (2020b) shows that $P_{subjective}(X_i) = P_{objective}(X_i)$ for all variables i if the DM's subjective DAG is *perfect*, that is, does not involve any v-colliders. Intuitively, someone viewing the data through a perfect DAG does not neglect the correlation between two variables that jointly cause a third. Such an agent cannot be 'systematically fooled'—for instance made to think that the mean of any variable is higher than it actually is. Spiegler (2020b) additionally shows that if the DGP features a multivariate normal joint distribution of $X_1, \ldots, X_n$, then any subjective DAG the DM could fit to the data will imply the correct marginal distribution for each i, no matter whether it involves v-colliders or not.

The above results characterize *when* DMs with misspecified subjective DAGs will make incorrect predictions, but not *how large* their errors will be. Eliaz et al. (2020) answer the latter question in a multivariate normal setting by studying the extent to which fitting a misspecified model to data can distort estimates of pairwise correlations. The question is germane, for instance, to a DM seeking to interpret medical research about the effect of a particular behaviors, such as the consumption of table salt (X_1) on longevity (X_3). Because longevity is typically hard to observe, medical studies often rely on surrogate endpoints that are known to correlate with longevity, such as blood pressure (X_2). Eliaz et al. (2020) consider a DGP in which X_1 and X_3 are uncorrelated ($\rho_{1,3} = 0$) and ask what is the largest correlation that an analyst could infer by fitting the model $X_1 \to X_2 \to X_3$? They show that for suitable joint distributions of X_1, X_2, and X_3, fitting this chain can yield an inferred correlation as high as 0.5 despite the absence of any true correlation between X_1 and X_3. In practice,

such spurious inference can occur if researchers opportunistically select one from a range of plausible surrogate endpoints (*marker hacking*). If the researcher can select multiple intermediate markers to create arbitrarily long causal chains to be fit, the inferred correlation between nodes that are in fact uncorrelated can be arbitrarily close to 1.

Spiegler (2025) considers the costs of a different type of error, namely the inclusion of the wrong set of control variables when agents seek to learn the causal effect of their actions on payoff-relevant outcomes from observational data. He finds that these welfare costs from estimating causal effects with bad controls differ greatly depending on whether the data from which agents learn are given exogenously or are endogenously produced by the agent's own actions.

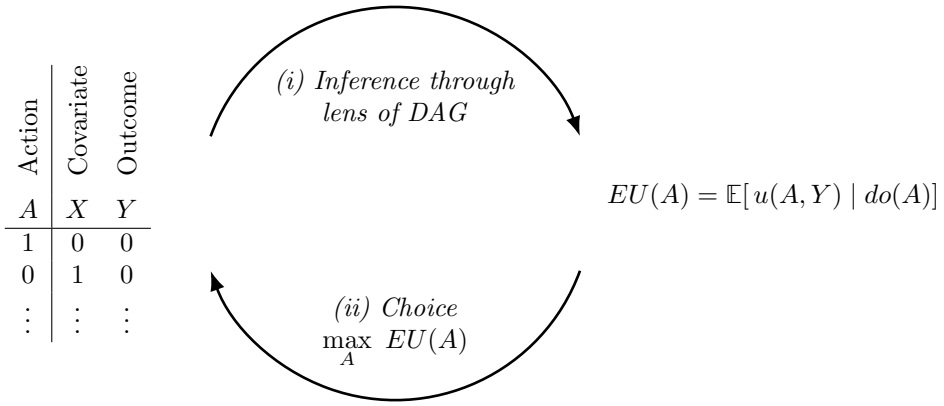
Choice with misspecified DAGs When a DM seeks to predict the effects of her own choices or other interventions with the causal system, Markov equivalence to the DGP is no longer sufficient to rule out errors. Instead, the DM's subjective DAG must be *identical* to that of the DGP. Otherwise, she will mispredict the effects of intervening on at least some variables for some parameterizations of the DGP. As in the case of non-interventional prediction, a DM with a misspecified DAG may mispredict only the effects of interventions that do not matter to her, perhaps because she does not have an opportunity to affect the relevant variables. For instance, the central bank can control interest rates but not government spending, and a manager may be able to influence the behavior of her subordinates but not that of her supervisors. Will some misspecifications be without consequence once we limit the set of nodes the agent can intervene on or cares about?

Spiegler (2016) provides a precise answer to this question. Suppose the DM can intervene on a single node A (his 'action') and has a utility function defined over some subset M of nodes. Even if the DM's DAG G is misspecified, she will correctly perceive the mapping from her action to all utility-relevant outcomes if and only if the set of action and utility-relevant nodes, $\{A\} \cup M$, forms an ancestral clique in some DAG that is Markov equivalent to G, and the DM does not observe the realizations of any variables before intervening.

Equilibrium effects The above results apply regardless of how the data used for learning is obtained. Particularly interesting effects arise when the DM herself generates the data as a byproduct of her actions. In such settings, the use of a misspecified model can induce beliefs that justify taking those very actions, even if they are objectively suboptimal.

Spiegler (2016)'s *Dieter's Dilemma* illustrates the key idea. It considers the case of a person who mistakes a (harmless) symptom of a disease for its cause. The person has access to a dietary supplement that masks the symptom but does not affect the disease, and she only cares about the disease. That is, the true causal structure is *Supplement* $\rightarrow$ *Symptom* $\leftarrow$ *Disease*, but the person believes *Supplement* $\rightarrow$ *Symptom* $\rightarrow$ *Disease*. When she does not consume the supplement, she observes a strong correlation between the disease and the symptom, which she mistakes for a causal

Figure 4: Personal equilibrium



Notes: Personal equilibrium requires that two conditions hold simultaneously. Condition 1: The joint distribution of all variables ('data') projected through subjective DAG justify the choices in condition 2. Condition 2: Choices produce the joint distribution of variables in condition 1.

effect of the symptom on the disease. She thus infers that the supplement must be effective in treating the disease, and increases her supplement intake. If she always consumes the supplement, the symptom is masked; she observes no correlation between the symptom and the disease. She infers the supplement must be ineffective, and decreases her consumption. If her intake is high, she will lower it, and if it is low, she will raise it. Therefore, there will be an equilibrium in which she consumes the supplement with some intermediate regularity. The equilibrium consumption depends both on the magnitudes of the actual causal effects and on the costs and benefits associated with each action and outcome, and thus allows for rich comparative statics. If the symptom is a less reliable indicator of the disease, for example, the person believes the supplement to be less effective and thus consumes less of it in equilibrium. Like all models we discuss in this subsection, the Dieter's Dilemma assumes that the decision maker follows her subjective causal model dogmatically; she does not process the information about how medication intake and disease covary to falsify her misspecified DAG.[12]

The Dieter's Dilemma illustrates the concept of *personal equilibrium* (Spiegler, 2016).[13] It concerns an agent who interprets data through a given causal model. Her actions determine the data she observes, which she in turn uses to update the parameters of her subjective causal model. Personal equilibrium characterizes a situation in which the distribution of actions taken by the agent generates precisely the data which, when viewed through the lens of her model, justify taking that action distribution (see Figure 4 for illustration). Formally, a probability distribution over actions is a

[12]The existence of this interior equilibrium requires additional assumptions. The cost of the supplement must be sufficiently low, for instance, for the decision maker to ever want to consume it.

[13]The concept has a similar flavor but is formally unrelated to personal equilibrium in the literature on reference-dependent preferences (Kőszegi and Rabin, 2006).

personal equilibrium if any action that the DM takes with positive probability maximizes her subjective expected utility according to the beliefs generated from fitting her subjective DAG to the data.

Personal equilibrium yields interesting applications that we discuss in Section 7. Besides generating novel insights, personal equilibrium also subsumes well-known behavioral frameworks as special cases. For example, Eyster and Rabin (2005)'s *cursed equilibrium* considers a specific form of misspecification in which players in a game neglect the information content of others' actions. In a market for lemons (Akerlof, 1970), a 'cursed' buyer ignores the fact that a seller has private information about the value of a car. This buyer can be described as processing the decision problem through a DAG that omits the link from the seller's knowledge about the car to the seller's willingness to accept. Personal equilibrium is also closely related to $S(K)$-*equilibrium* (Osborne and Rubinstein, 1998), *analogy-based equilibrium* (Jehiel, 2005), and *naive behavioral equilibrium* (Esponda, 2008). It is a special case of Esponda and Pouzo (2016)'s *Berk-Nash equilibrium*, which does not restrict misspecification to (unparametrized) DAGs but allows for general misspecified likelihood functions.

Even with misspecified DAGs, personal equilibrium effects do not always arise. When changing the frequency of the action leaves the agent's perceived mapping from actions to consequences unchanged—a property Spiegler (2016) calls *consequentialist rationality*—there is no feedback loop and the agent's optimization problem is standard. Spiegler (2016) shows that this property holds whenever every variable that the agent does not treat as a direct consequence of her action is, under the true DAG, independent of the action given the causes the agent does recognize. When consequentialist rationality is violated, Carbajal and Nachbar (2025) show that the resulting personal equilibrium effects are generic; they arise for all but a measure-zero set of distributions consistent with the true DAG.

While personal equilibrium considers agents whose models are structurally misspecified but correctly fit to the data, a complementary literature considers agents whose models are structurally correct but impose erroneous parametric restrictions. This approach has been used to explain phenomena such as misdirected learning, stereotyping, and bias substitution (Heidhues et al., 2018, 2025).

4.2 Empirical evidence

We next turn to the empirical question of how humans interpret data when they have beliefs about the structure of causal relationships between observable variables but not their strengths. To answer this question, we largely retain the empirical literature's focus on binary variables.

The literature has considered various definitions of the magnitude of a causal effect. The quantity known as ΔP (Jenkins and Ward, 1965) appears naturally in economic decision making. Consider an agent who decides whether to spend amount k to bring about a cause C (such as a treatment for an

illness, or an investment in education) to influence an outcome E. Doing so is optimal if

$$\sum_{e\in\{0,1\}} \big(P(E=e|\text{do}(C=1)) - P(E=e|\text{do}(C=0))\big)U(E=e) \geq k \tag{4}$$

where U denotes the agent's money-metric utility over E, which we assume is separable from the expense k. The measure of causal strength is simply defined as the change in the chance the effect obtains if the cause C is exogenously imposed present rather than absent, $\Delta P = P(E=1|\text{do}(C=1)) - P(E=1|\text{do}(C=0))$, which appears as the first factor in equation (4).[14]

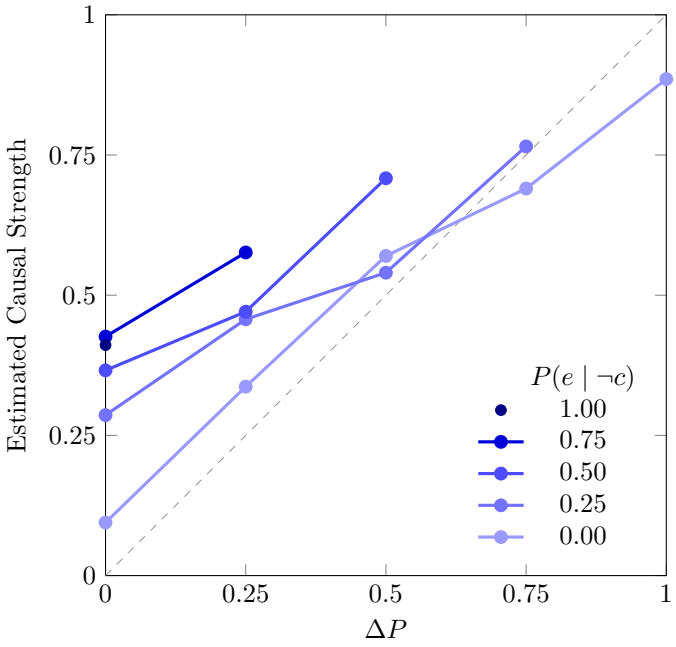
The key economic question regarding parameter learning is thus what individuals infer about ΔP in a given setting based on observable data and their beliefs about the causal structure of the DGP. We review the empirical research on this question in four steps. First, we consider learning when the data features causal effects, focusing on whether and how subjective probabilities are warped versions of objective probabilities. Second, we consider the important special case in which the DGP shows no causal effect, where subjects may nonetheless perceive non-zero causal effects. Both cases suggest that the apparently biased mapping from objective probabilities to subjective judgments arises in part because subjects reason as if unobserved background causes and hidden interactions are also at work. Third, we discuss research on the priors that people tend to use in parameter learning, which, if they are strong and data is finite, can bias their posteriors. Fourth, we discuss research that focuses on cases in which individuals' mental models may be structurally misspecified, both when they simply respond to observational data, and when they learn from the data they generate themselves, as in Spiegler's (2016) personal equilibrium.

Learning non-zero associations How well do subjects learn the magnitudes of existing causal associations? Ambuehl and Huang (2025)'s data provides an answer in a setting with binary variables. They regress subjects' reported beliefs about causal effects $P(Y|\text{do}(X=1,Z)) - P(Y|\text{do}(X=0,Z))$, where Z may be empty, on the causal effects observed in the data. They find substantial attenuation, with many subjects perceiving effects merely a quarter as large as the true association. By contrast, in a setting with continuous predictor and outcome variables—which entail much more information than binary realizations—Rehder (2025) finds that subjects' average estimates of the slope parameters linking variables are statistically indistinguishable from their actual values, though there is substantial heterogeneity across individuals.

While few papers put objective causal effects and their subjective counterparts on a directly comparable scale, a large literature in psychology on *elemental causal induction* studies how objective causal effects map onto Likert-scale ratings of 'causal strength' that are often not precisely defined

[14] While equation (4) refers to interventional probabilities, the original formulation of ΔP by Jenkins and Ward (1965) uses observational probabilities. The two cases coincide if C is exogenous.

Figure 5: Empirical mapping of objective probabilities into Likert-scale ratings of 'causal strength'



Notes: Results from Buehner et al. (2003). Treatments with constant $P(e \mid \neg c)$ in the data are connected by lines. The horizontal axis displays $\Delta P = P(e \mid c) - P(e \mid \neg c)$.

to subjects.[15] Hence, nonlinear or otherwise surprising mappings between such ratings and objective causal effects can arise either from biased inference or from interpreting the questions as eliciting something other than probabilities. Leaving this interpretational ambiguity unresolved, the literature finds that subjects' Likert-scale statements about causal strength strongly covary with ΔP, but systematically deviate from it depending on the probability with which the effect is present even when the cause is absent. Figure 5 shows a typical result. Buehner et al. (2003) argue that these judgments are well-described by the measure $q = \Delta P/(1 - P(E|\neg C))$, termed *Power PC* by Cheng (1997).

The fact that *Power PC* is not the relevant measure for economic decision making leads one to ask why it would describe subjects' statements. The seminal contribution of Griffiths and Tenenbaum (2005) argues that the reason is humans' tendency for causal rather than purely statistical thinking. They assume that individuals do not treat the effect's occurrence in the absence of the causes as just a statistical statement, but instead posit an alternative unobserved background cause B that could explain the effect. They show that if subjects assume the effect E is a common consequence of the cause C and the background cause B, linked through a noisy-OR specification, then the implied causal strength parameter for C is given by the Power PC measure. In a similar way, ΔP also corresponds

[15] For instance, Buehner et al. (2003) ask subjects to '*judge how strongly each vaccine prevented the disease related to the virus in question by giving a rating on a scale from 0 (the vaccine does not prevent the disease at all) to 100 (the vaccine prevents the disease every time).*'

to the implied strength parameter under a linear specification. It remains an open question whether humans think about causal strength in terms of Power PC even in economic decision making, when doing so can reduce utility, or whether Power PC merely describes their responses to survey questions.

Failure to learn the absence of associations While individuals sometimes perceive actual effects in attenuated form, they also show the opposite bias—perceiving correlations and causal effects where none exist (reviewed in Fiedler, 2016; Matute et al., 2015; Fiedler et al., 2022). In a typical experiment on such *illusory correlation*, subjects observe, for instance, data on 120 patients who differ in whether they received a medication and whether they recovered, with the frequencies shown in Panel A of Table 3. From this data, the average subject typically infers that the medication is effective, even though the probability of recovery is 80% regardless of medication intake. The effect appears when both the effect and the cause have skewed marginal distributions. Subjects behave as if common causes are associated with common outcomes, and rare causes with rare outcomes. When interpreted causally, illusory correlation has interesting implications for research on stereotyping, as one can see by relabeling the contingency table as Panel B of Table 3. Subjects typically infer from such data that minority members commit crimes more often than majority members, even though it shows no such association (Hamilton and Gifford, 1976).

Table 3: Illusory causation: Example contingency tables

A. Medical example

	Recovered	Did not recover
Took medication	80	20
Did not take medication	16	4

B. Stereotypes

	Did not commit crime	Committed crime
Majority member	80	20
Minority member	16	4

While illusory correlation may seem like a mistake, recent research argues that it is not necessarily irrational. One strand of research (Costello and Watts, 2019; Bott et al., 2021; Gershman and Cikara, 2023) shows how it can arise from Bayesian updating. Gershman and Cikara (2023), for instance, consider two groups with equal true means. If the assessor's Gaussian prior has a mean below the sample average and one group is observed more often, posterior means (precision-weighted averages of prior and data) will differ: the larger group's estimate is pulled closer to its empirical mean, yielding a higher posterior. A second strand attributes it to causal elaboration mechanisms. Following a similar argument to Griffiths and Tenenbaum's (2005) rationalization of the Power PC measure of causal strength, Ng et al. (2026) argue that people who observe the effect occurring in absence of the cause

may posit background causes that vary across conditions and thereby account for the effect instead. Indeed, in a medical scenario similar to that of Panel A of Table 3, their participants spontaneously assumed that patients who were not given the medication had stronger immunity or less severe illness than those who were given it. Positing this additional mechanism let them reconcile their belief in a positive causal effect with the statistical independence observed in the data.

The type of illusory causation just discussed arises purely from interpreting contingency tables, regardless of whether the subject merely observes data or takes actions in an attempt to influence outcomes. The *illusion of control* is a variant of this phenomenon that is likely not driven by illusory correlation. It occurs in situations such as die tossing, where subjects know that affecting outcomes is impossible, but still behave as if they could affect them (Langer, 1975; Langer and Roth, 1975). In cases where mere observation yields illusory causation, personal involvement does not strengthen perceived contingencies (Yarritu et al., 2014), though it does distort which variables subjects include in their causal models (Fan, 2024).

Priors Even if causal effect magnitudes can be learned correctly when given enough data, priors about parameters will affect beliefs when samples are finite. What priors do individuals tend to bring to new problems?

Lu et al. (2008) argue that human priors do not follow the uniform distribution across parameters that some papers assume for technical convenience (e.g., Griffiths and Tenenbaum, 2005). Instead, they argue that people 'prefer causal models that minimize the number of causes ... while maximizing the strength of each individual cause that is in fact potent.' Empirically, they showed that a formalization of such *strong and sparse* priors, when updated with data, better fits human posterior judgments than does updating uniform priors. To measure priors more directly, Yeung and Griffiths (2015) used an experimental design based on iterated learning. Under the assumptions of their model, this process defines a Gibbs sampler in which subjects' choice distributions would converge to their priors. Using this method, they confirmed that subjects' priors favored near-deterministic relationships between the main cause and the outcome. In contrast to Lu et al. (2008)'s proposal, however, their empirically identified priors showed no preference for the probability of the main cause C over alternative causes, and no competition between the probabilities of the main and background causes. The preference for (near-)determinism in people's priors appears to vary across individuals but stable for a given individual (Mayrhofer and Waldmann, 2015).

Learning with misspecified causal models Two different strands of empirical research focus on learning and choosing with misspecified models.

The first strand tests the main qualitative assumption of the misspecified models literature: do people with different subjective models interpret the same data differently? Because these studies seek to make a broad, qualitative point, they are not always set in the context of causal DAGs. Barron and

Fries (2023) show that externally provided interpretations for a series of ten data points (specifically, claims about a structural break in the DGP) significantly affected subjects' predictions, even when subjects knew that the information providers had incentives to bias their interpretation against the subjects' own interest. Kendall and Charles (2022) specifically focus on DAGs. They consider the setting of Eliaz and Spiegler (2020) in which subjects observe a table listing realizations of triples of binary variables A, X, Y. Their results show that subjects' choices could be biased by providing them with interpretations that suggested that the data were consistent with a chain $A \to X \to Y$ or a collider $A \to X \leftarrow Y$.[16] A caveat is that all interpretations included directives like '*So, choose* $A = 1$ *more often than* $A = 0$' which produced similar behavioral effects even when presented without any explanation of the data.

The second strand of research tests the specific predictions of the theory that individuals interpret data as if fitting a mixture of causal DAGs to it. Specifically, Ambuehl and Huang (2025) tested whether subjects' causal belief systems were rationalizable as arising from interpreting data through the lens of a causal DAG or a mixture of DAGs. They also tested the predictions of Spiegler (2016)'s personal equilibrium concept, conducting three types of tests. First, for the Dieter's Dilemma, they derive a theoretical result showing that the only subjective DAG that can rationalize believing in the ability to affect the outcome is the chain, $A \to X \to Y$. Any subject who interprets the data through this DAG, however, must also believe that X blocks the path from A to Y. Ambuehl and Huang (2025) found that this prediction was empirically violated. Second, acknowledging that model uncertainty may cause subjects to behave as if by interpreting the data through a mixture of DAGs, they identified a set of restrictions that beliefs consistent with this theory must satisfy. Their data also rejected these restrictions. Third, they tested the comparative statics predictions of Personal Equilibrium, for instance by varying the parameters of the DGP to increase the equilibrium action frequency, or by altering the link function to produce multiple equilibria. The data directionally matched the predicted comparative statics, but these effects were substantially attenuated. This attenuation may reflect factors such as choice stochasticity, which could more generally weaken the model's predictions across settings in which Spiegler (2016)'s model applies.

5 Structure learning

In the previous sections, we treated the structure of the causal model as given, focusing on how agents reason within that structure and estimate its parameters. However, this presupposes that agents hold beliefs about which structure describes the causal relationships between variables in the first place. How should idealized agents acquire causal structure, and how do humans actually do so?

[16] These interpretations take the form '$X = 1$ *only if* $A = 1$. *Further, when* $X = 1$, $Y = 1$ *is always true. So, choose* $A = 1$ *more often than* $A = 0$' and '*If* $X = 0$, *it always holds that* $Y = 0$ *if* $A = 1$. *To counteract this, choose* $A = 0$ *more often than* $A = 1$ *so that* Y *can be 1 even if* $X = 0$.'

A natural question is why agents would seek to infer structure rather than simply estimating parameters in a single, fully connected model. The answer is that structure constrains inference, and thereby makes learning more efficient. By ruling out many possible relationships among variables, learners can draw stronger conclusions from limited data than a completely unstructured approach would permit.

In this section, we first review normative accounts of structure learning (Section 5.1), including constraint-based and Bayesian frameworks, as well as hierarchical formulations that allow structure to be shared across contexts. We then examine empirical evidence (Section 5.2) on how accurately individuals infer causal structure, the conditions under which they succeed or fail, and the processes by which they update beliefs about structure.

5.1 Theoretical concepts

5.1.1 Constraint-based learning

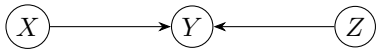
A frequentist approach to inference about causal structure is *constraint-based* learning (Spirtes et al., 2000). The main idea is to use the set of conditional independence relations implied by each causal DAG to narrow down the set of candidate structures, ideally to the Markov equivalence class of the DGP's causal structure. For example, the chain and fork structures in Panels A and C of Figure 2, relabeled to $X \rightarrow Y \rightarrow Z$ and $X \leftarrow Y \rightarrow Z$, respectively, imply that X and Z are independent conditional on Y. By contrast, the v-collider of Panel B, relabeled to $X \rightarrow Y \leftarrow Z$, implies that X and Z are unconditionally independent. Hence, if a dataset of observations of X, Y, and Z reveals independence between X and Z conditional on Y but no other independence relations, this observation is consistent with the chain $X \rightarrow Y \rightarrow Z$ and the fork $X \leftarrow Y \rightarrow Z$ (as well as the reverse chain $X \leftarrow Y \leftarrow Z$), which form a Markov equivalence class, but not with other DAGs. If, on the other hand, the dataset shows independence of X and Z, but dependence between all other pairs of variables, both conditionally and unconditionally, the only compatible three-node DAG is the v-collider of Panel B. In practice, independence is asserted based on null hypothesis significance tests. In settings with more than a handful of variables, the vast number of possible DAGs requires specialized algorithms to discover the Markov equivalence class of the DAG consistent with the list of conditional independencies.

5.1.2 Bayesian structure learning

The cognitive science literature more commonly models structure learning in a Bayesian framework. This approach offers several advantages, some of which make it particularly appealing as a model of human judgment. First, Bayesian methods quantify uncertainty by assigning posterior probabilities to alternative causal structures. Constraint-based methods, in contrast, simply output a set of DAGs deemed consistent with the data. Second, the Bayesian framework extends readily to hierarchical

Figure 6: Bayesian structure learning example.

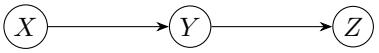
Collider



$P(X{=}1) = P(Z{=}1) = 0.50$

$$P(Y{=}1 \mid X, Z) = \begin{cases} 0.25 & X{=}0,\ Z{=}0 \\ 0.75 & \text{otherwise} \end{cases}$$

Chain



$P(X{=}1) = 0.50$

$$P(Y{=}1 \mid X) = \begin{cases} 0.75 & X{=}1 \\ 0.25 & X{=}0 \end{cases}$$

$$P(Z{=}1 \mid Y) = \begin{cases} 0.75 & Y{=}1 \\ 0.25 & Y{=}0 \end{cases}$$

Prior: $P(\textit{collider}) = P(\textit{chain}) = 0.5$

Data. Observed: $(X, Y, Z) = (0, 1, 1)$. Likelihood of this data point:

$$P\big((X, Y, Z) = (0, 1, 1)\big) = \begin{cases} 0.19 & \text{if DAG} = \text{Collider} \\ 0.09 & \text{if DAG} = \text{Chain} \end{cases}$$

Posterior:

$$\begin{aligned} P(\text{DAG} = \text{Collider} \mid \text{data}) &= \frac{P\big((X, Y, Z) = (0, 1, 1) \mid \text{Collider}\big)\, P(\text{DAG} = \text{Collider})}{P(\text{data})} \\ &= \frac{0.19 \times 0.5}{0.19 \times 0.5 + 0.09 \times 0.5} = 0.68. \end{aligned}$$

Notes: In this example, the learner has no uncertainty about the parameters of the DAG. In general, the learner faces a distribution over parametrizations of the DAGs. She uses data to update not only about the likelihoods of the structures, but also about the parameters within each structure.

settings in which different environments share regularities (the same physical laws, for instance, govern both a bike rolling downhill and a door swinging shut). This makes it well-suited to capture how people learn abstract higher-level patterns and generalize across contexts. Third, it can handle data of different types—combining, for example, historical observations with verbal explanations. Fourth, it naturally accommodates individual differences in prior beliefs, allowing agents to assign different posteriors to DAGs that are observationally similar. Fifth, Bayesian models have inspired approximate process accounts meant to better incorporate the cognitive constraints of human learners.

In principle, the mechanics of Bayesian structure learning are straightforward: calculate the probability of the observed data for each structure, then use them in Bayes' law to update beliefs about structures. Figure 6 shows an example. A learner entertains two candidate models: a collider DAG and a chain DAG. For simplicity, suppose the learner knows the parameters of each structure, and assigns equal prior probability to the two DAGs. How should she update her beliefs upon observing the data point $(X, Y, Z) = (0, 1, 1)$? To answer this question, recall that a parametrized DAG implies a probability distribution over all possible outcomes. Hence, we can calculate each candidate DAG's

probability of producing observation $(X, Y, Z) = (0, 1, 1)$. These probabilities, along with the prior probability of each model, determine the posterior probability of each model according to Bayes' law. In this specific example, the observation increases the learner's belief that the DGP is the collider DAG from 0.5 to 0.68.

The above example illustrates the general principle—a simple restatement of Bayes' law—that the posterior probability of a model given the data is proportional to the probability of the data given the model,

$$P(\text{model}|\text{data}) \propto P(\text{data}|\text{model})P(\text{model}). \tag{5}$$

If the parameters of each structure are also uncertain, we calculate the likelihood that a given structure produced the data by averaging over the distribution of possible parameter values:

$$P(\text{data}|\text{model}) = \sum_{parameters} P(\text{data}|\text{model, parameters})P(\text{parameters}|\text{model}).$$

The two factors on the right-hand side of equation (5) highlight important properties of structure learning. The likelihood, $P(\text{data}|\text{model})$, captures an intuitive feature of how people learn structure; upon observing data, we increase our weight on the structures that are more likely to have produced it and decrease our weight on those that are less likely to have produced it. More surprisingly, it embeds a preference for simplicity in structure learning via the so-called *Bayesian Occam's Razor*. Consider a simple model and a more complex one, and suppose both can explain a given set of observations similarly well. More complex models typically admit a wider range of parameter values, many of which fit the observed data poorly, but the marginal likelihood averages over all parameter values consistent with a model. As a result, the complex model is penalized more heavily, and beliefs update more strongly toward the simpler model (MacKay, 2003).[17]

Priors impose ex ante tendencies on structure learning, and can be formulated either directly over the DAGs themselves via $P(model)$, or indirectly over the parameters (Lu et al., 2008) or functional form (Lucas and Griffiths, 2010) conditional on a DAG. By encoding assumptions about the typical strength, sparsity, or determinism of causal relationships, prior beliefs influence which structures appear most plausible. For instance, beyond the tendency to favor simpler models induced by Bayesian Occam's Razor, a preference for simpler or more complex structures may stem from learners' priors over DAGs, as when those priors favor coarser partitions of variables (Kemp et al., 2010b).

Once an agent has drawn inferences about the likelihood of multiple structures, how does she make predictions or choose between actions? There are two common approaches (Koller and Friedman, 2009). *Model selection* entails picking the single most probable structure (e.g., the *maximum a posteriori*, or MAP, estimate) and choosing the action implied as optimal by that structure. In

[17] While the Bayesian Occam's Razor has rarely been the explicit focus of work on structure learning, human cognition shows consistency with some of the hallmarks of Bayesian Occam's Razor in other settings (Blanchard et al., 2018).

model averaging, by contrast, learners weight predictions about the effects of interventions by each structure's posterior probability and choose the action that yields the best outcome according to these weighted predictions.

While the data in the example above comprises a vector of binary outcomes, the Bayesian framework can accommodate and integrate data of any kind. For instance, information is often communicated through verbal explanations (Sloman et al., 2009; Lombrozo and Vasilyeva, 2017; Meder and Mayrhofer, 2017), which can be cast as probabilistic evidence within a Bayesian model (Nam et al., 2023), if one provides an explicit stochastic mapping from causal structures to natural language statements. While straightforward in principle, implementation is challenging in practice because this mapping must operate over the vast space of possible utterances (Bar-Asher Siegal, 2026).

5.1.3 Bayesian Hierarchies and the Blessing of Abstraction

The Bayesian approach to structure learning raises the question of where priors about causal structure come from. In general, they may reflect past experience, background assumptions, or abstract knowledge carried over from other domains. *Hierarchical Bayesian models* provide an influential formal account of how priors can be learned from experience across settings.

Hierarchical Bayesian models extend ordinary Bayesian inference by allowing priors themselves to depend on higher-level variables that are also learned from data (Griffiths and Tenenbaum, 2009). In the case of structure learning, this means treating the causal structure of a particular system as a draw from a higher-level distribution over structures, which is itself inferred from experience. At the lower level, the learner infers the causal structure governing the system at hand—such as how the success of a specific firm depends on factors like location and product quality. At the higher level, the learner infers regularities about what kinds of causal graphs are typical across systems in a domain—for example, learning that firms generally succeed due to multiple interacting causes rather than a single dominant factor. These hierarchies can be extended arbitrarily—at a still higher level, one might infer that social and economic systems like firms tend to be described by densely connected graphs, whereas many physical systems are represented by more sparsely connected graphs.

Kemp et al. (2010a) illustrate these ideas in a setting involving biomedical data. Their hierarchical Bayesian model receives a set of entities (e.g., "lung cancer," "ammonia," "fever") and observed relational predicates between them (e.g., "causes," "affects"), with no labels indicating what kind of entity anything is. At the higher level, the model learns a framework theory specifying categories such as chemicals, diseases, symptoms, organisms, as well as the laws that connect them (e.g., chemicals cause diseases, diseases produce symptoms). At the lower level, these categories and laws constrain inferences about specific relationships between individual entities, such as whether asbestos causes lung cancer. Even though neither level is given in advance; their model successfully learns the categories, laws, and specific relational structure from raw data. While Kemp et al. (2010a) is about discovering relational systems of concepts in general, Kemp et al. (2010b) study the same hierarchical logic in an

explicitly DAG-based setting. At the lower level, their model learns the causal properties of individual objects, while at the higher level, it learns which kinds of objects tend to produce which kinds of effects. Because the model captures abstract regularities shared by different objects, it can infer the properties of a new object from as little as a single observation. More fundamentally, Goodman et al. (2011) show that even the abstract principles of the causal DAG framework, such as the fact that dependence is directed and that interventions are exogenous, can be learned at a higher level, while specific causal graphs are learned at a lower level.

The hierarchical modeling approach thus addresses the provenance of priors by treating them as the product of Bayesian inference at higher levels. It also explains how learning transfers between contexts. Observations in one case shape beliefs about others insofar as they inform higher-level inferences shared across contexts. This feature of hierarchical models can give rise to what is known as the *Blessing of Abstraction* (Goodman et al., 2011; Gershman, 2017b). In contrast to the *Curse of Dimensionality* which makes unconstrained learning over high-dimensional distributions impractical, hierarchical inference allows learners to pool information across environments and extract abstract regularities supported by a larger effective sample size than any single instance could provide. Once learned, these abstractions narrow down the hypothesis space in new contexts and thereby accelerate inference. Hence, the blessing of abstraction makes hierarchical Bayesian inference exceptionally potent, and helps explain how people can learn so much from what seems like so little data.

5.2 Empirical evidence

The empirical literature on structure learning is vast. We first review evidence on the question of whether and how well subjects can learn structure at all (Section 5.2.1), and whether their beliefs display the typical hallmarks of constraint-based (Section 5.2.2) or Bayesian learning (Section 5.2.3). Because structure learning can be computationally intensive, we then consider some processes by which humans are thought to approach Bayesian solutions, as well as deviations that stem from cognitive limits (Section 5.2.4).

5.2.1 How well do people learn structure?

Broadly speaking, empirical research in cognitive science shows that people are able to learn causal structure from data, though perhaps more slowly and less accurately than ideal learners (Rottman, 2017). For example, participants in Coenen et al. (2015) were able to select the correct three-node DAG from two options 87% of the time when allowed to freely intervene on the DAG (see also Coenen et al., 2019). When given finite observations, various studies find that participant performance matches Bayesian benchmarks (e.g., Nam et al., 2023). Even children as young as four years old draw correct causal inferences from observing interventions in simple cases (Gopnik et al., 2001, 2004). Related evidence from economics points in the same direction. Frechette et al. (2024) presented subjects with

datasets of triples of binary variables. Subjects studied these datasets knowing that they would later have to predict variables based on observations of other variables in the same system using only the notes they had taken, without access to the original dataset. Subjects' largely successful predictions indicated considerable efficacy in extracting the DAG structure of the DGP.

In addition to such conscious, deliberate learning, structure inference occurs at an automatic, perceptual level (reviewed in Shams and Beierholm, 2022). The *ventriloquist illusion*, a canonical example of *sensory cue integration* (Körding et al., 2007; Sato et al., 2007), illustrates this point. Observers must infer where a voice comes from when they both see a puppet and hear 'its' voice. The perceptual system appears to consider two possibilities: that the voice comes from the puppet's *seen* location (common cause), or from somewhere else (separate causes). When what is heard is near what is seen, the common cause model gets more weight, and the voice is perceived as coming from where the puppet is *seen*. As the mismatch increases, the separate cause model gets more weight, and the perceived source shifts back toward where the voice is *heard*. This non-monotonic pattern in the distance between the perceived and actual location of the voice as a function of the distance between the puppet's seen location and the location of the voice is a signature of Bayesian model averaging.

At the same time, a large body of work documents substantial failures of causal structure learning. For instance, the subjects in Lagnado and Sloman (2004) performed no better than chance when asked to pick the correct causal structure from five options after observing realizations from a three-variable probabilistic system for 50 trials. Similarly, the subjects in White (2006) performed near chance levels when seeking to infer structure from observational data as written sentences describing patterns of co-occurrence between populations of species in a nature reserve. Fernbach and Sloman (2009) also document poor structure learning from observation. Indeed, performance can vary greatly within the same studies (e.g., Lagnado and Sloman, 2002; Steyvers et al., 2003).

What explains this variation in the success of causal structure learning? Performance is typically higher when subjects can actively intervene (Lagnado and Sloman, 2002; Coenen et al., 2015), when information about temporal order or delay is available (Lagnado and Sloman, 2006; Bramley et al., 2018), when strong domain priors or hierarchical structure guide inference (Griffiths and Tenenbaum, 2009; Kemp et al., 2010b), when the system is deterministic or low-noise (Mayrhofer and Waldmann, 2016; Rothe et al., 2018; Frechette et al., 2024), and when each effect has few causes (Rothe et al., 2018). In short, success is more likely when the available information is richer and the underlying structure is simpler or less noisy.

5.2.2 Constraint-based learning

While cognitive science largely focuses on settings suited to Bayesian structure learning, some economic applications are better captured by constraint-based learning. Consider agents who choose among externally proposed models (for instance, because they arise in public policy debates), possibly by comparing them to statistical summaries of historical data. Ambuehl and Thysen (2024) study

this case in a stylized four-variable environment where subjects could inspect aggregate statistics of the DGP—that is, any conditional or unconditional correlation among pairs or triplets of variables. Subjects were shown pairs of candidate models, each recommending a different investment level, and chose which to follow. Since payoffs depended on the chosen investment and the true DGP, they faced an incentive to identify whether the correctly specified model was in the choice set, and, if so, select it. The models were described in natural language (e.g., 'X directly affects Y,' 'X influences Y indirectly through Z,' or 'X is a symptom of Z') and illustrated with DAGs. They find that individuals displayed a remarkable ability to discard misspecified models; three fifths of their subjects did so consistently. Success in this task required not only intuiting the relevant correlational implications but also finding which of 18 possible charts displaying empirical data would help check these implications. Subjects' performance stemmed from the possibility of relying on qualitative (i.e., whether model-predicted correlations coincide with those observed in the data) rather than quantitative inference (i.e., how strongly investments and payoffs covary in the data). In fact, subjects rarely even viewed the data necessary for quantitative inference. Moreover, failures to detect misspecified models reflected limited ability rather than limited motivation, as evidenced by the fact that tripling the monetary stakes (up to $90 in some cases) had no effect.

5.2.3 Hallmarks of Bayesian structure learning

Other applications may be better described by Bayesian structure learning, especially those in which agents learn from the piecemeal arrival of individual observations. Beyond the fact that people routinely draw causal inferences from samples too small for standard statistical tests to detect reliable dependencies (Gopnik et al., 2001), several lines of evidence indicate that human learning bears several of the distinctive features of Bayesian structure inference. Among these are findings that people simultaneously entertain and update weights on multiple causal structures, that their inferences reflect prior beliefs about the domain, and that these priors can themselves be updated with experience.

Simultaneous inference over competing structures Evidence from Gershman (2017a) suggests that learners do not commit to a single causal structure, but instead maintain and update beliefs over multiple candidates. In the experiment, participants learned relationships among three variables: the cue, the context, and the outcome. In one condition, context was irrelevant; in a another, it could affect the outcome probability independently; in a third, it modulated the effect of the cue on the outcome. Subjects then made predictions for combinations of cues and contexts, including in cases where the context variable took values not seen before. When the context was irrelevant, participants inferred that the cue signaled the outcome regardless of context; when the context had an independent effect, participants inferred that its effect on the outcome did not depend on the cue; but subjects were reluctant to generalize when cue and context both mattered together. This pattern is parsimoniously explained by a Bayesian model that assigns probabilities to the candidate structures based on the

observations. Related evidence comes from sequential decision-making tasks studied by Acuña and Schrater (2010), where participants' behavior was consistent with maintaining a posterior over causal structures rather than committing to a single one. In these tasks, participants had to learn whether rewards across options were independent or coupled. When the options were likely independent, subjects explored to gather information about each option, whereas when they were likely coupled, they exploited more aggressively because data on one option was informative about the other. When the evidence was ambiguous, behavior lay between these extremes, changing smoothly with the inferred likelihood of each structure. Though these studies do not directly measure beliefs over structures (cf. Steyvers et al., 2003), they are consistent with the idea that learners simultaneously entertain multiple causal models and update their relative plausibility over time.[18]

Sensitivity to prior information Judgments of causal structure can be influenced in a graded manner by prior beliefs about the presence, strength, or functional form of causal links. Griffiths and Tenenbaum (2009) review various strands of evidence for this claim and demonstrate how judgments quantitatively match the implications of Bayesian structure learning. For example, manipulating the base rate of causation affects beliefs about whether objects cause a detector to activate, even with the same subsequent evidence. Other findings point to the role of informative priors over causal strength or functional form. In the classic task of elemental causal induction (see Section 4.2), where people learn about a single cause-effect relationship, Lu et al. (2008) show that models with priors favoring sparse and strong causes explain both strength judgments and structure judgments better than models with uninformative priors. Likewise, Lucas and Griffiths (2010) demonstrate how priors over functional form influence judgments of structure by exposing participants to causal systems governed by either a conjunctive form (where both causes must be present to yield an effect) or a disjunctive form (where either alone suffices). Under a conjunctive form, repeated failures by a single candidate cause are not so informative, since it may need a partner cause; under a disjunctive form, the same failures count strongly against causation. Accordingly, participants given conjunctive training largely ignored such failures when later evaluating a novel candidate cause, as predicted by a Bayesian model.

Hierarchical Bayesian inference Kemp et al. (2010b) study whether people use what they learned in earlier, similar situations to form expectations in new ones. In their experiments, participants saw objects that affected a machine, where some objects had the same causal effects and could therefore be grouped together, even though these groupings were not directly stated. When asked to evaluate a new object, participants did not rely only on the limited data about that object. Instead, they interpreted this information in light of patterns they had seen before. As a result, the same sparse evidence (sometimes just visible features) led to different conclusions depending on what participants had previously learned. For example, after seeing that earlier objects fell into two groups with different

[18] These papers belong to a much larger literature in cognitive science on the importance of causal structure in reinforcement learning (Gershman, 2017c).

effects, participants tended to assign a new object to one of these groups, so the same observation was interpreted differently depending on which grouping seemed more likely. Moreover, these expectations were flexible. When new observations repeatedly contradicted the earlier grouping, participants updated their beliefs and were willing to revise the grouping itself. A hierarchical Bayesian model captures this behavior by allowing people to update both which group an object likely belongs to and what effect objects in each group tend to have.

5.2.4 Bounded rationality in structure learning

While the above evidence suggests that Bayesian structure learning provides a good as-if model of human causal cognition, it may be too computationally demanding to cognitively implement in full detail. Indeed, people rely on coarse, schematic representations that capture broad causal patterns while leaving many mechanistic details unspecified (Keil, 2003). The *illusion of explanatory depth* reveals that this incompleteness often goes unnoticed (Rozenblit and Keil, 2002). More generally, people appear willing to tolerate or even prefer compressed causal models when additional detail seems unnecessary for the task at hand (Kinney and Lombrozo, 2024).

How do humans actually learn causal structure? A prominent view in the cognitive science literature proposes that they infer larger causal structures through sequences of local updates rather than by globally evaluating all candidate structures. The evidence in Fernbach and Sloman (2009), for instance, suggests that people evaluate individual causal links in a piecemeal fashion, and focus on evidence that is immediately available and easy to process. Specifically, when a variable X is intervened on, people tend to treat the resulting pattern of activations on that trial as evidence for direct causal links from X to each variable that becomes active, and they build toward a more global understanding by accumulating such local inferences. Consistent with this account, experimental participants made systematic errors when inferring chains $X \to Y \to Z$, frequently believing in a spurious, direct $X \to Z$ connection (Frechette et al., 2024, document the same pattern).

Purely local computations raise the concern that they need not result in a globally coherent belief system or correspond to a normatively well-grounded learning process. Bramley et al. (2017) addresses this issue by proposing a form of local learning based on Gibbs sampling, a method for approximate inference. They call this the 'Neurath's Ship' algorithm, alluding to a metaphor attributed to Otto van Neurath: '*We [theorists] are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support.*' In the algorithm, the learner maintains a single current DAG hypothesis and, after each new observation, performs a small number of local edits: it selects an edge and updates its state (absent, $i \to j$, or $j \to i$) by sampling conditional on the current graph and recent evidence. Updates are restricted to preserve acyclicity, and belief updating is both local and conservative, favoring small changes to the current hypothesis. In addition, past evidence is partially forgotten or downweighted, inducing a recency bias. This model matches two key

empirical facts. First, people's belief revisions about structure tend to be sticky; they rarely make large jumps between structures. Second, revisions exhibit a recency bias, consistent with the model's assumption that older evidence is discounted.

Why does continued exposure to the DGP not necessarily correct misspecified beliefs? A number of papers speak to this issue, though not framed specifically in terms of DAGs. Schwartzstein (2014) shows theoretically that if agents do not initially attend to a variable, they may never learn its relevance and may instead become more convinced that the variables they do track are the important ones; Hanna et al. (2014) observe this phenomenon among seaweed farmers. Gagnon-Bartsch et al. (2023) extend this logic to characterize when agents with misspecified models will ever realize their model is wrong. A broad class of related *learning traps* has been also studied in cognitive science (Rich and Gureckis, 2018).

Even when the relevant models are known, people may be unable to fully compute their implications. In the experiment of Musolff and Zimmermann (2025), participants faced uncertainty about which of two explicitly specified models determined outcomes. The experimenters manipulated complexity by varying whether participants were given the already precomputed outcomes under each model or had to instead calculate the outcomes themselves before combining the models. Faced with the latter burden, participants calculated outcomes based only on the more likely model, despite otherwise continuing to acknowledge model uncertainty. In that sense, complexity can induce a shift from model averaging to model selection, not by changing which models people consider plausible but by making it harder to integrate them.

6 Methodology: Inferring and eliciting beliefs about causality

Most of the research discussed so far examines beliefs or judgments about specific causal relationships rather than complete belief systems. This raises two related questions. First, are individuals' choices or elicited beliefs consistent with a coherent subjective causal model? Evidence at the level of isolated comparative statics may appear consistent with the causal DAG framework even if judgments violate global structural restrictions (such as coherence between beliefs about direct, indirect, and total effects). Second, how can subjective causal structure be measured empirically, and to what extent do different elicitation methods capture the underlying belief system rather than artifacts of the measurement procedure? The next two subsections tackle these questions in turn.

6.1 Rationalizability

Theory Theoretical work characterizes when a DM's choices can be represented as expected utility maximization under a single subjective causal model, and how that model can be identified. Approaches differ in the scope of feasible interventions as well as in the nature of the choice sets they consider and thus in the identifying variation they generate. Halpern and Piermont (2024) and

Schenone (2024) consider a setting in which the DM can freely intervene on any variable in the network. Halpern and Piermont (2024) recover a subjective DAG from preferences over interventions themselves (building on Pearl's machinery); Schenone (2024) recovers them from intervention-act pairs, where interventions alter the subsequent evaluation of acts (building on Savage's machinery).

By contrast, Ellis and Thysen (2025) emphasize that economic agents typically control only a small subset of nodes in the causal system determining their utility, and that utility may in turn depend on few nodes. An employee, for instance, may only be able to control their effort and may just care about remuneration. Hence, they consider a DM who can affect merely a single node (the 'action') and whose utility depends on a single node (the 'outcome'), though their results extend to the case of multiple outcomes if these are all directly causally related to one another. This restriction makes identification more demanding, as the analyst just observes stochastic choices over actions rather than preferences over arbitrary interventions. Nonetheless, they show that if the utility function is known, the stochastic choice rule follows a logit form, and the subjective DAG contains no v-colliders, then choices from a single dataset identify the subjective DAG up to a relation they call *prediction-equivalence*. Loosely, DAGs are prediction-equivalent if they share the same confounders (which they define as variables that the DM believes can influence both her action and another variable) and the same shortest directed paths from the action and each confounder to the outcome. The main contribution of Ellis and Thysen (2025) concerns the general case in which the utility function and the payoff node are unknown, and the data from which the DM learns may be affected by her own actions.

Are empirical beliefs DAG-rationalizable? The decision-theoretic work does not yet lend itself to empirical application because it either requires having subjects make an impractically large number of choices (Halpern and Piermont, 2024; Schenone, 2024) or makes the identifying assumption that individuals make no mistakes when fitting models to data (Ellis and Thysen, 2025), in addition to the expected utility assumption that empirical choices generally violate. To empirically test whether individuals' beliefs are consistent with the causal DAG framework, Ambuehl and Huang (2025) devise the following method. They define a *Causal Belief System* $\mathcal{B}$ as the list of all interventional probability statements defined over a set of random variables and elicit subjects' beliefs about all these probabilities. Letting $\mathcal{P}^D$ denote the corresponding interventional probability statements obtained by fitting DAG D to the data, they call a subject's belief system *DAG-mixture rationalizable* if there exists a probability distribution q over all DAGs such that $\mathcal{B} = \sum_D q_D \mathcal{P}^D$ and *DAG rationalizable* if $q_D = 1$ for some D. In other words, this test assesses whether a person's elicited beliefs could potentially stem from some weighted average of possible DAGs. Because the number of possible DAGs grows extremely quickly in the number of nodes, one might suspect that nearly any belief system is DAG-mixture rationalizable. However, this is not the case, because the predictions generated by different DAGs are highly constrained and overlap substantially, causing the vectors $\mathcal{P}^D$ to occupy a

low-dimensional subspace of the space of all possible belief systems. The set of belief systems that can be represented as a mixture of DAGs is therefore much smaller than the set of all logically possible belief systems. Empirically, in a setting with relatively noisy DGPs, Ambuehl and Huang (2025) find that subjects' elicited belief systems are substantially closer to the best fitting DAG mixture than to a uniformly random benchmark—about three times closer in terms of Euclidian distance, both with and without allowing for probability weighting (Kahneman and Tversky, 1979). At the same time, while the causal DAG framework has explanatory power, a substantial portion of variation remains unexplained.

6.2 Elicitation

Studying causal beliefs often requires measuring them directly. Existing methods to elicit causal structure fall into three broad classes: numbers (quantitative judgments), words (verbal descriptions), and pictures (graphical depictions). The approaches differ both conceptually and in how well they perform empirically.

Conceptually, numerical elicitation has the advantage of imposing the weakest assumptions. Unlike graphical elicitation, it does not presume that human belief systems conform to the Causal DAGs formalism. Unlike verbal elicitation, it does not require the analyst to make decisions about the mapping from words to numerical quantities or DAGs. Yet, the elicitation of all quantities required to describe a causal network is large and quickly becomes unmanageable as the number of nodes grows. Moreover, requiring subjects to consider all possible relations also risks pointing out causal paths they might otherwise have ignored.

Verbal elicitation has the advantage of most closely matching how people naturally talk about causation. Due to the unclear mapping between verbal statements and formal models, however, the resulting representations may be incomplete or ambiguous. Whether this incompleteness reflects an artifact of the elicitation method or a genuinely underspecified belief system remains an open question.

Graphical elicitation has the advantage of speed, completeness, and an ostensibly clear link to the formal framework. Conceptually, these advantages come at the price of assuming that individuals' causal belief systems are consistent with the causal DAGs formalism. It also needs modification to allow the expression of uncertainty over models. Even when these assumptions are satisfied, it is still unclear whether subjects interpret their own graphical depictions the same way an analyst would. If, for instance, a subject draws a collider DAG, does this mean that they are aware, and mean to convey, that conditioning on the collider will cause an apparent correlation between the source nodes? Moreover, the key information in a DAG is not the arrows that are present but those that are absent—arrows merely represent the possibility, but not the necessity of causal influence, whereas the absence of an arrow definitely precludes influence. It is unclear whether subjects appreciate this asymmetry when constructing diagrams.

Empirically, there is scant evidence that compares the approaches. One exception is Ambuehl and Huang (2025), who elicit subjects' entire set of interventional and marginal beliefs over three binary variables after 100 rounds of interaction, and also ask them to depict their beliefs about causal relations graphically. They find that the list elicitation predicts choices well, but that the graphical depictions predict neither choices nor beliefs elicited in the list format.

Other studies also compare list elicitation to graphical depictions, use benchmarks other than predicting consequential behavior. Tatlidil et al. (2025) find—though under extremely strong identifying assumptions—that list elicitation produces elicited beliefs closer to the true DGP in a physical framing. They conjecture that this outperformance is due to the fact that list elicitation forces subjects to consider all potential causal connections. Green et al. (2003) finds similar results in a social psychological context.

Verbal elicitation has been studied in a line of research that seeks to quantify the information conveyed by statements such as 'A killed B' (strong) or 'A caused B's death' (less strong) (Rose et al., 2021; Beller and Gerstenberg, 2025), including through Large Language Models (Priniski et al., 2023). This literature leaves open several interesting questions about how verbally elicited beliefs relate to consequential choices, objective causal relations in the world, and alternative elicitation methods.

Within-study consistency is suggestive about the types of inferences that can and cannot be drawn from verbal statements.[19] Andre et al. (2024) ask respondents *'Which factors do you think caused the increase in the inflation rate? Please respond in full sentences.'* Two human research assistants then codify these responses in the form of DAGs. Andre et al. (2024) find moderate inter-rater reliability on the level of individual links between nodes: if one coder assigns a causal connection between two specific factors, there is a 77% chance that the other coder does so as well. Researchers seeking to understand subjective causal models, however, often need a more demanding benchmark than the coincidence of links, as it is three-node substructures that capture the key qualitative properties of causal DAGs. The rate at which the two coders map statements to exactly the same DAGs is 51%. This number drops further once we consider DAGs that are not simply a list of factors which each influence the outcome (star-networks): if one coder maps the verbal statement to a DAG that involves at least one indirect effect, the chance that the other coder maps it to the same DAG is 21%. The modest performance of the hand-labeling method is consistent with Yang et al. (2022) who also document its brittleness and report poor out-of-sample predictive performance. It is an open question whether this brittleness represents the ability of human research assistants (which natural language processing methods might solve), or inherent conceptual issues with the objective of mapping each respondent's verbal statement to a single DAG.[20]

[19] A body of research has more generally studied the mapping between numerical probabilities and their verbal expression (Wallsten and Budescu, 1983).

[20] While Frechette et al. (2024) ask subjects to write notes to their future selves to help with prediction tasks in binary variable DAGs, they do not extract subjects' DAGs from these notes.

These results illustrate that the choice of method must be informed by the researchers' aims. For instance, verbal elicitation may suffice if the goal is to identify which variables enter a subjective causal belief system, but more formal elicitation methods may be necessary to study beliefs about the structure of relations between these variables.

7 Applications

Research on causal cognition has been used to inform political economy (Section 7.1), microeconomics (Section 7.2), macroeconomics and finance (Section 7.3), and business (Section 7.4). Here, we briefly discuss a selection of research that applies the causal DAG framework in each of these fields.

7.1 Political economy

Some of the most exciting applications of causal DAGs in economics concern political economy, where they provide a way to represent narratives as explicit causal models. This approach has been used to explore phenomena such as the persistence of competing narratives in public discourse, cycles of populism, and belief polarization.

The theory of narrative competition proposed by Eliaz and Spiegler (2020) shows how conflicting causal narratives can persist simultaneously. It yields striking implications: false narratives can survive in equilibrium, and it is possible to predict the prevalence of narratives in a way that permits comparative statics. These results rest on the key assumption that individuals adopt whichever model promises the higher utility when fit to the data, without regard for its accuracy. To illustrate the logic, consider the debate over whether mask-wearing causally reduces COVID-19 transmission. Suppose there are two narratives, formalized as subjective causal models. Model 1 correctly states that increased masking reduces COVID-19 transmission. Model 2 holds that masking has no effect on transmission. To see how, first consider the case in which most individuals adopt Model 1. In this case, most will wear masks, and case counts will be low. According to Model 1, this situation can be maintained only by continued masking, which has a hassle cost. Model 2 promises higher utility because it predicts case counts will remain low even if everyone stops masking and thus saves the hassle cost. Therefore, individuals will begin to adopt Model 2 and cease masking. Second, consider the case in which most individuals adopt Model 2. In this case, only few individuals will wear masks, and case counts will be high. According to Model 2, masking cannot change this situation. Model 1 promises higher utility because it predicts that case counts can be lowered at a small hassle cost. Individuals will thus flock to Model 1. Overall, therefore, the more popular one model is, the more attractive the other becomes; narratives 'feed off each other.' Accordingly, no model, including the true one, can survive alone; narrative multiplicity is a necessary outcome. Equilibrium is achieved when the models are adopted with a frequency that makes them equally attractive. Comparative

statics predictions arise from the fact that parameters such as the hassle costs of mask-wearing and the strength of causal effects in the DGP determine the equilibrium frequencies.

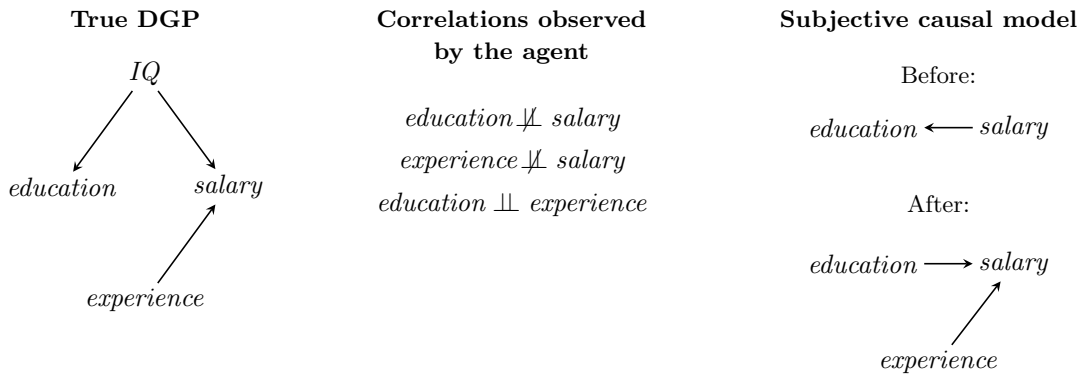
While Eliaz and Spiegler (2020)'s model is insightful, its empirical accuracy remains contested. Angrisani et al. (2024) demonstrate that it can rationalize the dynamics of Americans' beliefs about the effectiveness of preventive behaviors during the COVID-19 pandemic. Yet, Ambuehl and Thysen (2024)'s direct experimental tests of its key assumption find little empirical support. The reason is that for a DM who views narrative adoption as a choice under ambiguity, Eliaz and Spiegler (2020)'s key assumption is equivalent to maximizing best-case outcomes. Most of their subjects instead maximize worst-case outcomes, consistent with a general tendency observed in decisions under ambiguity (Trautmann and Van De Kuilen, 2015).

Other work on subjective causal models in political economy does not assume that voters adopt narratives at will, but are instead married to a subjective causal model. Their turnout to the voting booth depends on the extent to which their model promises that their favored party would produce better outcomes than the competition. Levy et al. (2022) show how this assumption, when placed in a dynamic political economy framework, can yield cycles of populism in which a populist party and a mainstream party take turns governing. Their mainstream party has the correct model of the world and chooses optimal policy, whereas the populist party's model omits some relevant variables. Turn-taking in governing arises due to the following process. When the populists observe outcomes generated under the mainstream party's policies, omitted-variable bias leads them to overestimate the effectiveness of the policies they consider relevant and to advocate more extreme policies. This raises their turnout and brings them to power. Once in power, their own policy outcomes gradually lead beliefs to correct, reducing their perceived advantage and eventually returning power to the mainstream party, at which point the cycle begins anew.

Also assuming that the strength of causal beliefs determines turnout, Eliaz et al. (2022) find that political coalitions will often find it optimal to spin false narratives that attribute desirable outcomes to the exclusion of particular social groups ('scapegoating') rather than the policies that were actually responsible.

Other research uses the framework of causal DAGs to understand belief polarization, a phenomenon in which the same information leads people with different priors to update in opposite directions. Collider DAGs can naturally produce polarization through the logic of explaining away (following the spirit of Botvinik-Nezer et al., 2023). Suppose a candidate's victory depends on two causes: actual majority support and election fraud. Consider two voters with different priors about majority support; one is confident that the majority favors the candidate, whereas the other is confident the majority opposes her. Both believe that fraud, if it occurs, would favor the candidate. Observing that the candidate won causes them to update their beliefs about fraud in opposite directions. The first voter, who expected the win based on majority support, sees fraud as less necessary to explain the outcome, while the second, who did not expect the win, finds fraud more plausible. Similar arguments can be

Figure 7: Causal persuasion (Burkovskaya and Starkov, 2026).



constructed to provide rational explanations for many classical studies of belief polarization (Jern et al., 2014).

7.2 Microeconomics

Principal-agent theory The causal DAG framework has yielded new insights in principal-agent theory in settings such as contracting and persuasion.

Schumacher and Thysen (2022) study contracting when the agent evaluates actions through a misspecified causal model. Their main example features a marketer who seeks to maximize sales, and must decide whether to make cold calls. In their setting, cold calling makes customers more informed but damages the firm's reputation, and both of these factors in turn affect sales. However, the agent's causal DAG is misspecified: she neglects the reputation channel and thus thinks cold calling pays off more than it really does. The principal can therefore induce an action that would otherwise be too costly, because the agent overestimates its effectiveness.

A larger number of papers study persuasion. Some of the most active research on persuasion through suggested interpretations relies on the *model persuasion* framework by Schwartzstein and Sunderam (2021) that conceives of subjective models as likelihood functions but does not explicitly involve causality. In their baseline model, a principal and an agent both observe the same state-dependent dataset. The principal would like to induce an action that differs from the agent's preferred choice in that state. She does so by proposing an interpretation of the data in the form of a likelihood function, which the agent adopts if it fits the data better than his default model. Two central results are that, with finite datasets, persuasion is generally possible, and that agents tend to follow overfitted models.[21] Barron and Fries (2023) test this model experimentally and show that human subjects

[21] Related research extends this framework in several directions: to organizational settings (Schwartzstein and Sunderam, 2022); to environments in which the principal does not observe the specific data realization before deciding which model to send (Aina, 2021); to settings that incorporate elements of Bayesian Persuasion (Ichihashi and Meng, 2021; Jain, 2023); and to experimental tests of whether subjects select the best-fitting model (Aina and Schneider, 2025).

indeed manage to influence other's actions by proposing interpretations even when both observe the same data. Further expanding this line of research using the toolbox of causal DAGs is promising for two reasons. First, whenever there are multiple variables, likelihood functions are impractical to convey and may be replaced with causal explanations that can be modeled as causal DAGs. Second, assessing the quantitative fit of any model is difficult for humans. But causal stories in the form of DAGs allow subjects to evaluate models based on qualitative fit (e.g., whether the data display the presence or absence of correlations implied by the story), which many individuals manage to do well (Ambuehl and Thysen, 2024).

Burkovskaya and Starkov (2026) take a step in this direction. They analyze a setting in which a principal who knows the DGP can influence the beliefs of an agent who is unaware of some variables. The agent interprets data by forming subjective causal DAGs in the manner of a constraint-based algorithm (see Section 5.1). They show that making the agent aware of neglected variables can lead him to reverse his beliefs about the direction of causality between other variables. Their main example, illustrated in Figure 7, centers on a worker who believes that *education* $\leftarrow$ *salary* when in fact both *education* and *salary* are independent effects of a third variable *IQ* of which the worker is unaware. The agent is also unaware of the variable *experience* that influences *salary*. The principal seeks to reverse the agent's beliefs about the direction of causation between *education* and *salary*. She achieves this by pointing out that *experience* is correlated with *salary* but independent of *education*. The worker infers that the only DAG consistent with this pattern of correlations is *education* $\rightarrow$ *salary* $\leftarrow$ *experience*. Burkovskaya and Starkov (2026)'s general analysis of such persuasion strategies finds that persuasion is only possible when the receiver's initial model conflicts with the true causal structure, and that making an agent believe in the existence of a link is often easier (in the sense that it requires disclosing just one or two variables) than making them believe that a link is absent (which requires disclosing every common cause). It is an open empirical question whether disclosing correlations between neglected variables causes agents to reassess the direction of causality in the way the model assumes, or instead prompts them to search for additional variables that reconcile the newly revealed correlations with their initial beliefs (Papakonstantinou et al., 2024).

Behavioral economics Many applications of the causal DAG framework intersect with behavioral economics, which frequently studies the impact of biased beliefs and imperfect inference. We have already covered formal connections to other theories in behavioral economics (see Section 4.1.2) and contributions to topic areas that typically fall within its purview, such as self-confirming false beliefs (section 4.1.2), stereotyping (section 4.2), and failures to detect misspecification (section 5.2.4). Causal DAGs also bear on other issues of relevance, including moral decision making (through the attribution of blame and credit; Lagnado et al., 2013; Gerstenberg et al., 2018; Engl, 2022), categorical thinking (Rehder and Hastie, 2001; Rehder, 2017a,b), and confirmation bias and motivated reasoning (Gershman, 2019; Caddick and Rottman, 2021; Pilgrim et al., 2024).

7.3 Macroeconomics and finance

Macroeconomics is a fruitful field for applying the causal DAG framework, since expectation formation plays a central role in many macroeconomic questions. Applications have taken both theoretical and empirical directions.

Theoretically, Spiegler (2020b) revisits the classical question of whether the Phillips curve can be systematically exploited by a central bank, but under the assumption that the private sector forms beliefs by fitting a a DAG to observed data, rather than holding rational expectations. In his leading example, the private sector endorses a 'classical' narrative, wherein it perceives output as a cause of inflation rather than the reverse, and excludes expectations from its model entirely, as did monetary theory prior to Phelps (1967) and Friedman (1968). Because this DAG treats the central bank's action and output as independent co-causes of inflation without linking them directly, it contains a v-collider. In this case the central bank can systematically push inflation above private-sector expectations. However, when the private sector's DAG is perfect (that is, when it does not contain a v-collider), or when the objective distribution is multivariate normal, no such exploitation is possible (see Section 4.1.2). This result puts the historical debate surrounding the Lucas critique into a new perspective: non-exploitability results do not necessitate rational expectations; they survive under a much broader class of causal models.

While Spiegler (2020b) asks whether a strategic outside party can exploit an agent's causal misperceptions, Spiegler (2022) shifts the focus to what happens when the agent with the wrong causal model is the DM, so that her behavior shapes the very data from which she draws faulty inferences. The leading macroeconomic application is again the Phillips curve. Following Sargent (1999) and Cho and Kasa (2015), the private sector's subjective model inverts the Phillips curve, meaning that it treats employment as an independent explanatory variable for inflation, whereas in the true model causality runs from inflation to employment. The true DGP also includes fundamentals that directly affect inflation but not employment. When the private sector bases its inflation forecasts on this inverted Phillips curve, forecast errors arise when anticipated inflation has real effects, as follows.[22] To forecast inflation, the private sector conditions on fundamentals and employment and then averages over employment using its unconditional distribution, as if employment were an exogenous shock unrelated to fundamentals. But when anticipated inflation has real effects, employment is in fact correlated with fundamentals through its dependence on inflation. By averaging over employment as if this correlation did not exist, the private sector dilutes the information that fundamentals carry about inflation, pulling forecasts toward the unconditional mean and making them too rigid. The extent of this dilution depends on the private sector's own forecasting behavior. A more rigid forecast entails less variation in anticipated inflation, which changes how much employment covaries with

[22] Under the new-classical assumption that only unanticipated inflation matters, no systematic forecast errors arise. The reason is that in this case, employment reduces to the forecast error plus noise, which is by construction uncorrelated with the predictors in the private sector's inverted Phillips curve; treating it as exogenous is therefore harmless.

fundamentals and thereby how much the exogeneity assumption distorts the next round of inference. The personal equilibrium is the forecasting rule at which the rigidity implied by this distortion exactly matches the rigidity that generated the data to which the wrong model was fit. Hence, in equilibrium, forecasts underrespond to changes in fundamentals.

Empirical work on causal cognition in macroeconomics shows how misspecified beliefs about the causal structure of the macroeconomy cause demand for counterproductive policies. Binetti et al. (2024), for instance, find that 60% of Americans erroneously believe that high interest rates cause high inflation and support rate cuts to fight it. This finding parallels Andre et al. (2022), who show more broadly that laypeople's intuitions about macroeconomic mechanisms differ systematically from those of experts. Regarding the relation between interest rates and inflation, they find, as do Binetti et al. (2024), that a majority of households predict that an interest rate hike increases inflation whereas most experts predict the opposite. They show that households disproportionately think of a 'cost channel' in which firms pass higher borrowing costs on as higher prices, whereas experts invoke the textbook demand channel. While theoretical models often employ the representative agent paradigm, Andre et al. (2022) also document substantial heterogeneity in households' subjective models of the macroeconomy. Relatedly, Andre et al. (2024) study narratives about the 2021-2022 U.S. inflation surge and find that households' narratives are simpler and more fragmented than those of experts and differ systematically in terms of the variables they consider to be causes of inflation. Households focus disproportionately on supply-side factors such as supply-chain disruptions and the energy crisis, and often feature politically loaded explanation such as government incompetence or corporate greed which are entirely absent from expert narratives.

The implications of misspecified subjective mental models have also been studied in finance, partly following the influential presidential address of Shiller (2017). While this work does not explicitly deploy the causal DAG framework, research by Molavi et al. (2024), for instance, is closely related. They consider an asset-pricing environment in which fundamentals are driven by a hidden n-factor model. Agents can observe the realized fundamentals but not the underlying factors. Agents seek to understand these fundamental processes by estimating hidden-factor models. Due to complexity constraints, these models feature fewer factors than are present in the DGP. This model generates both predictable returns and predictable forecast errors. When applied to foreign exchange markets, it produces two classic violations of uncovered interest rate parity, namely the forward discount puzzle and the predictability reversal puzzle.

7.4 Business

Research on causal cognition has furthered our understanding of both marketing and strategic management by clarifying how consumers and managers perceive causal relationships and act on those perceptions.

Marketing One stream of research in marketing investigates which kinds of claims about a product's causal mechanism are convincing to consumers. Fernbach et al. (2013) find that more reflective individuals (as measured by the cognitive reflection test) prefer more detailed mechanistic explanations of how products work. Bharti and Sussman (2025) examine product claims that invoke causal chains, and show that people prefer products when all links in that chain are directionally consistent (e.g., a supplement increases alertness by increasing hormone production) compared to when the chain involves both positive and negative links (e.g., the supplement increases alertness by reducing hormone production). They argue that directionally consistent chains are easier to process, in turn raising perceptions of product efficacy.

Kupor and Laurin (2020) and Daniels and Kupor (2023) study the link between the perceived probability and magnitude of product outcomes. The former paper shows that people expect more likely outcomes to be larger; for example, when participants were told that tomato consumption had a higher probability of increasing metabolism, they expected the resulting increase in metabolism to be greater. The latter paper documents the converse pattern, wherein people infer that relationships with larger perceived magnitudes are more likely to reflect causal effects. Both effects are consistent with the interpretation that people believe that strong underlying causes increase both the probability and the magnitude of effects.

Other research examines *causal dilution*, the sometimes-observed phenomenon that the perceived strength of a cause declines with the number of effects it has. Sussman and Oppenheimer (2020) found causal dilution when additional effects were beneficial—subjects judged products to be less effective for their main purpose—but the opposite when the effects were harmful. Stephan et al. (2023) suggest that the lack of dilution in the latter case may have resulted from a kind of causal elaboration in which participants spontaneously added causal links among the negative outcomes. For example, if a shaving cream causes both ingrown hairs and skin irritation, people may infer that the ingrown hairs also worsen the irritation. Controlling for this mechanism yielded causal dilution in both positive and negative conditions. Zhang and Paolacci (2026) point out that when causal structure is uncertain, the opposite of causal dilution can occur. Accordingly, experimental participants inferred more strongly that drinking tea improves bone health when they also learned that tea consumption is correlated with better heart health. The authors argue that when people simultaneously evaluate both structure and strength, additional correlations make the focal outcome seem more plausibly affected by the same cause.

Causal DAGs can shed light other phenomena in marketing such as *backfire effects* (Bhui and Gershman, 2020). These arise when information that appears favorable is interpreted negatively. A product's star rating in online retail, for instance, could reflect both genuine quality and fake reviews. Under suitable parametric assumptions, Bayesian observers treat ratings that are high but still plausible as evidence of product quality, but interpret implausibly high ratings as a sign of fake. Inferred product quality can thus decline when signals are considered too good to be true. Experimental partic-

ipants in such a setting became more likely to suspect fraud as ratings increased, especially when prior information made fraud seem more likely. Stronger fraud inferences were associated with a weaker or even negative effect on purchase attitudes, and in many cases the same individuals responded both positively and negatively to higher ratings, depending on that prior information.

Strategic management A research program in strategy termed *Bayesian Entrepreneurship* (Agrawal et al., eds, 2026) draws on insights from Bayesian learning to understand and improve entrepreneurship. One of its primary tenets emphasizes the heterogeneous priors that individuals hold. DAGs afford us a means of characterizing this heterogeneity beyond simple quantitative differences in degrees of belief, capturing instead more fundamental differences in how individuals see the world. Camuffo et al. (2024a), for instance, argue that entrepreneurs' novel theories—formalized in their framework as DAGs—are a key source of competitive advantage. They illustrate this hypothesis with the case of the eyewear manufacturer Luxottica, which rose to prominence by reconceptualizing eyewear as a part of the fashion industry. This work falls within a broader literature on the *theory-based view of strategy* (Felin and Zenger, 2017), with Ehrig et al. (2024) focusing specifically on causal reasoning in the tradition of Pearl (2009). Empirically, Camuffo et al. (2024b) show in a series of large randomized controlled trials that this theory-driven, evidence-based approach (e.g., Felin et al., 2021) improves entrepreneurial decision making.

8 Conclusion and further reading

The history of economics features many examples of mathematical frameworks that had been fruitfully applied in other sciences for decades before economists started using them, including calculus, the theory of random networks, and information theory. We believe that causal networks offer another such valuable addition to the economist's toolkit.

Given its multi-decade interdisciplinary history, it is impossible to cover all aspects of causal cognition in an article-length review. Among the many fascinating questions that exceed the scope of this review are the following.

What are the entities that are linked by causal networks? This is a question of *causal ontology* (see Griffiths and Tenenbaum, 2009). It is relevant, for example, to differences in economic intuition between experts (who think about GDP, unemployment, and the money supply as separate entities) and laypeople (who may collapse these into a single entity called 'the economy'). Related work in cognitive science examines how individuals identify what belongs in a causal system and how their concepts change with experience (Goodman et al., 2007; Rottman et al., 2012; Goldwater and Gentner, 2015).

Where should the boundaries of a causal network be drawn, given that in principle every variable in the universe may be related to every other? This is the *causal frame problem* (Shanahan, 2016). One

approach in cognitive science is to treat it as a cost-benefit analysis where the greater accuracy from larger and more detailed models is weighed against their computational cost (Icard and Goodman, 2015). Of course, this cannot be a complete solution since it still presupposes a way of figuring out which parts of a potentially unbounded causal model are worth considering in the first place.

What leads people to single out some causes rather than others? Why, for instance, do we call an arsonist the cause of a fire but not the atmosphere, even though both the lit match and the presence of oxygen were necessary for the house to burn? This is the domain of research on *causal responsibility* judgments (e.g., Lagnado et al., 2013), which shows that people hold remarkably consistent intuitions as to which of multiple events in a causal model is the most important cause of an outcome (Morris et al., 2018). The economic model by Alexander and Gilboa (2023) hypothesizes that people regard x as a cause of y to the extent that accounting for x reduces y's Kolmogorov complexity. In economic policy applications, one might conjecture that such a systematic focus may direct attention toward specific policies (such as attributing economic development to new road construction) while neglecting the necessity of others (such as maintaining water infrastructure).

How do people make inferences when we cannot observe all variables in a causal network? Reasoning about such *hidden causes* has been studied extensively in cognitive science (e.g., Hagmayer and Waldmann, 2007; Rottman et al., 2011; Kushnir et al., 2010). In economics, Samuelson and Steiner (2025) study this issue within a broader framework based on variational autoencoders.

How do people draw parallels between different causal systems? We have already seen Bayesian models which incorporate forms of abstraction permitting agents to generalize their knowledge across settings. Generalization can extend further still as exemplified by analogical reasoning, where systems may be similar only in their relational structure even though the variables themselves are entirely different (Holyoak et al., 2010). We speculate that future research in this vein might be informative about lay economic reasoning. For example, insights from applications of causal DAGs to intuitive physics (reviewed in Gerstenberg and Tenenbaum, 2017) could shed light on how individuals (mis)apply concepts such as momentum or velocity in their mental models of economic phenomena.

The causal DAG framework has also been applied to illuminate fields outside of economics and cognitive science. In legal scholarship (reviewed in Lagnado and Gerstenberg, 2017) the framework facilitates progress on difficult conceptual questions about how to formalize notions like the *probability of necessity and sufficiency* (Pearl, 2009) that clarify cases such as the arsonist example above. Pertaining to public health, Powell et al. (2023) show that a detailed understanding of people's causal theories surrounding vaccines can help with the development of more effective informational interventions. Recent research also applies paradigms from cognitive science to artificial intelligence, in tests of whether Large Language Models reason causally like humans (Dettki et al., 2025), representing a return to the framework's computational roots.

From Aristotle to AI, the study of causation traces a venerable intellectual history to which economists are no strangers. Yet economics has only begun to engage with the flourishing litera-

ture on how ordinary people think about cause and effect, one that currently links fields as diverse as psychology, philosophy, and computer science. The research reviewed in these pages offers a sophisticated suite of theoretical and empirical tools that mesh well with existing approaches. We see myriad opportunities to understand how people construct their worldviews, to explore what leads them to diverge, and ultimately to close the gap between economic theories and economic life. The returns to this enterprise are, in our mental models, substantial.

References

Acuña, Daniel E. and Paul Schrater, "Structure learning in human sequential decision-making," *PLoS Computational Biology*, 2010, *6* (12), e1001003.

Agrawal, Ajay, Arnaldo Camuffo, Alfonso Gambardella, Erin L. Scott, Scott Stern, and Joshua Gans, eds, *Bayesian Entrepreneurship*, Cambridge, MA: MIT Press, 2026. Hardcover.

Aina, Chiara, "Tailored stories," *Unpublished*, 2021.

_ **and Florian Schneider**, "Weighting competing models," 2025.

Akerlof, George A., "The market for "lemons": Quality uncertainty and the market mechanism," *Quarterly Journal of Economics*, 1970, *84* (3), 488–500.

Alexander, Yotam and Itzhak Gilboa, "Subjective causality," *Revue économique*, 2023, *74* (4), 619–633.

Ambuehl, Sandro and Heidi Thysen, "Choosing between causal interpretations: An experimental study," *Unpublished*, 2024.

_ **and Jindi Huang**, "Models of causal misperceptions: Experimental tests," *Unpublished*, 2025.

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart, "Subjective models of the macroeconomy: Evidence from experts and representative samples," *Review of Economic Studies*, 2022, *89* (6), 2958–2991.

_ **, Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart**, "Narratives about the macroeconomy," Technical Report, SAFE Working Paper 2024.

Angrisani, Marco, Anya Samek, and Ricardo Serrano-Padial, "Competing narratives in action: An empirical analysis of model adoption dynamics," Technical Report, National Bureau of Economic Research 2024.

Angrist, Joshua D., "Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records," *American Economic Review*, 1990, pp. 313–336.

Barron, Kai and Tilman Fries, "Narrative persuasion," *WZB Discussion Paper*, 2023.

Beller, Ari and Tobias Gerstenberg, "Causation, meaning, and communication," *Psychological Review*, 2025.

Benjamin, Daniel J., "Errors in probabilistic reasoning and judgment biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2019, *2*, 69–186.

Bharti, Soaham and Abigail B. Sussman, "Consumers prefer products that work using directionally consistent causal chains," *Journal of Consumer Research*, 2025, *52* (2), 308–330.

Bhui, Rahul and Samuel J Gershman, "Paradoxical effects of persuasive messages," *Decision*, 2020, *7* (4), 239.

Binetti, Alberto, Francesco Nuzzi, and Stefanie Stantcheva, "People's understanding of inflation," *Journal of Monetary Economics*, 2024, *148*, 103652.

Blanchard, Thomas, Tania Lombrozo, and Shaun Nichols, "Bayesian Occam's razor is a razor of the people," *Cognitive Science*, 2018, *42* (4), 1345–1359.

Bohren, J. Aislinn and Daniel N. Hauser, "Misspecified models in learning and games," *Annual Review of Economics*, 2024, *17*.

Bott, Franziska M., David Kellen, and Karl Christoph Klauer, "Normative accounts of illusory correlations," *Psychological Review*, 2021, *128* (5), 856.

Botvinik-Nezer, Rotem, Matt Jones, and Tor D. Wager, "A belief systems analysis of fraud beliefs following the 2020 US election," *Nature Human Behaviour*, 2023, *7* (7), 1106–1119.

Bramley, Neil R., Peter Dayan, Thomas L. Griffiths, and David A. Lagnado, "Formalizing Neurath's ship: Approximate algorithms for online causal learning," *Psychological Review*, 2017, *124* (3), 301–338.

_ **, Tobias Gerstenberg, Ralf Mayrhofer, and David A. Lagnado**, "Time in causal structure learning," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2018, *44*, 1880–1910.

Buehner, Marc J., Patricia W. Cheng, and Deborah Clifford, "From covariation to causation: A test of the assumption of causal power," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2003, *29* (6), 1119–1140.

Burkovskaya, Anastasia and Egor Starkov, "Causal persuasion," January 2026. Working paper.

Caddick, Zachary A. and Benjamin M. Rottman, "Motivated reasoning in an explore–exploit task," *Cognitive Science*, 2021, *45* (8), e13018.

Camerer, Colin F. and Robin M. Hogarth, "The effects of financial incentives in experiments: A review and capital-labor-production framework," *Journal of Risk and Uncertainty*, 1999, *19* (1–3), 7–42.

Camuffo, Arnaldo, Alfonso Gambardella, and Andrea Pignataro, "Theory-driven strategic management decisions," *Strategy Science*, 2024, *9* (4), 382–396.

_ , _ , Danilo Messinese, Elena Novelli, Emilio Paolucci, and Chiara Spina, "A scientific approach to entrepreneurial decision-making: Large-scale replication and extension," *Strategic Management Journal*, 2024, *45* (6), 1209–1237.

Carbajal, Juan Carlos and John Nachbar, "Robust personal equilibrium effects in misspecified causal models," *Journal of Mathematical Economics*, 2025, *118*, 103127.

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

Cho, In-Koo and Kenneth Kasa, "Learning and model validation," *Review of Economic Studies*, 2015, *82* (1), 45–82.

Coenen, Anna, Bob Rehder, and Todd M. Gureckis, "Strategies to intervene on causal systems are adaptively selected," *Cognitive Psychology*, 2015, *79*, 102–133.

_ , Jonathan D. Nelson, and Todd M. Gureckis, "Asking the right questions about the psychology of human inquiry: Nine open challenges," *Psychonomic Bulletin & Review*, 2019, *26* (5), 1548–1587.

Costello, Fintan and Paul Watts, "The rationality of illusory correlation.," *Psychological Review*, 2019, *126* (3), 437.

Daniels, David P. and Daniella Kupor, "The magnitude heuristic: Larger differences increase perceived causality," *Journal of Consumer Research*, 2023, *49* (6), 1140–1163.

Davis, Zachary J. and Bob Rehder, "A process model of causal reasoning," *Cognitive Science*, 2020, *44* (5), e12839.

Dettki, Hanna M., Brenden M. Lake, Charley M. Wu, and Bob Rehder, "Do large language models reason causally like us? Even better?," *arXiv preprint arXiv:2502.10215*, 2025.

Ehrig, Timo, Teppo Felin, and Todd Zenger, "Causal reasoning and the scientific entrepreneur: Beyond Bayes," 2024.

Eliaz, Kfir and Ran Spiegler, "A model of competing narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

_ , _ , and Yair Weiss, "Cheating with models," *American Economic Review: Insights*, 2020.

_ , Simone Galperti, and Ran Spiegler, "False narratives and political mobilization," *arXiv preprint arXiv:2206.12621*, 2022.

Ellis, Andrew and Heidi Christina Thysen, "Subjective causality in choice," *arXiv preprint arXiv:2106.05957*, 2025.

Engl, Florian, "A theory of causal responsibility attribution," *Available at SSRN 2932769*, 2022.

Enke, Benjamin, Thomas Graeber, Ryan Oprea, and Jeffrey Yang, "Behavioral attenuation," *NBER Working Paper*, 2024.

_ , Uri Gneezy, Brian Hall, David Martin, Vadim Nelidov, Theo Offerman, and Jeroen Van De Ven, "Cognitive biases: Mistakes or missing stakes?," *Review of Economics and Statistics*, 2023, *105* (4), 818–832.

Esponda, Ignacio, "Behavioral equilibrium in economies with adverse selection," *American Economic Review*, 2008, *98* (4), 1269–1291.

_ **and Demian Pouzo**, "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 2016, *84* (3), 1093–1130.

Eyster, Erik and Matthew Rabin, "Cursed equilibrium," *Econometrica*, 2005, *73* (5), 1623–1672.

Fan, Tony Q., "Choice-induced misspecified mental models," 2024.

Felin, Teppo, Alfonso Gambardella, and Todd Zenger, "Value lab: A tool for entrepreneurial strategy," *Management and Business Review*, 2021, *1* (2), 68–76.

_ **and Todd R Zenger**, "The theory-based view: Economic actors as theorists," *Strategy Science*, 2017, *2* (4), 258–271.

Fernbach, Philip M, Adam Darlow, and Steven A Sloman, "Neglect of alternative causes in predictive but not diagnostic reasoning," *Psychological Science*, 2010, *21* (3), 329–336.

_ **,** _ **, and** _ , "Asymmetries in predictive and diagnostic reasoning.," *Journal of Experimental Psychology: General*, 2011, *140* (2), 168.

Fernbach, Philip M. and Steven A. Sloman, "Causal learning with local computations," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2009, *35* (3), 678–693.

Fernbach, Philip M, Steven A Sloman, Robert St Louis, and Julia N Shube, "Explanation fiends and foes: How mechanistic detail determines understanding and preference," *Journal of Consumer Research*, 2013, *39* (5), 1115–1131.

Fiedler, Klaus, "Illusory correlation," in "Cognitive Illusions. Intriguing Phenomena in Thinking, Judgment and Memory," 2 ed., Routledge, 2016.

_ **, Karolin Salmen, and Florian Ermark**, "Illusory correlation," in "Cognitive Illusions," Routledge, 2022, pp. 92–107.

Frechette, Guillaume, Emanuel Vespa, and Sevgi Yuksel, "Extracting models from data sets: An experiment," *Unpublished*, 2024.

Friedman, Milton, "The role of monetary policy," *American Economic Review*, 1968, *58* (1), 1–17.

Gagnon-Bartsch, Tristan, Matthew Rabin, and Joshua Schwartzstein, *Channeled attention and stable errors*, Harvard Business School Boston, 2023.

Gershman, Samuel J., "Context-dependent learning and causal structure," *Psychonomic Bulletin & Review*, 2017, *24* (2), 557–565.

_ , "On the blessing of abstraction," *Quarterly Journal of Experimental Psychology*, 2017, *70* (3), 361–365.

_ , "Reinforcement Learning and Causal Models," in Michael R. Waldmann, ed., *The Oxford Handbook of Causal Reasoning*, Oxford University Press, 2017, chapter 17, pp. 295–306.

_ , "How to never be wrong," *Psychonomic Bulletin & Review*, 2019, *26* (1), 13–28.

_ **and Mina Cikara**, "Structure learning principles of stereotype change," *Psychonomic Bulletin & Review*, 2023, *30* (4), 1273–1293.

Gerstenberg, Tobias and Joshua B. Tenenbaum, "Intuitive theories," in Michael R. Waldmann, ed., *The Oxford Handbook of Causal Reasoning*, New York: Oxford University Press, 2017, pp. 515–548.

_ **, Tomer D. Ullman, Jonas Nagel, Max Kleiman-Weiner, David A. Lagnado, and Joshua B. Tenenbaum**, "Lucky or clever? From expectations to responsibility judgments," *Cognition*, 2018, *177*, 122–141.

Glymour, Clark N., *The mind's arrows: Bayes nets and graphical causal models in psychology*, MIT Press, 2001.

Goldwater, Micah B. and Dedre Gentner, "On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories," *Cognition*, 2015, *137*, 137–153.

Goodman, Noah D., Tomer D. Ullman, and Joshua B. Tenenbaum, "Learning a theory of causality," *Psychological Review*, 2011, *118* (1), 110–119.

_ **, Vikash K. Mansinghka, and Joshua B. Tenenbaum**, "Learning grounded causal models," in "Proceedings of the 29th Annual Conference of the Cognitive Science Society" 2007, pp. 305–310.

Gopnik, Alison, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir, and David Danks, "A theory of causal learning in children: Causal maps and Bayes nets," *Psychological Review*, 2004, *111* (1), 3–32.

_ , David M. Sobel, Laura E. Schulz, and Clark Glymour, "Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation.," *Developmental Psychology*, 2001, *37* (5), 620.

Green, D.W., S.J. Muncer, T. Heffernan, and I.C. McManus, "Eliciting and representing the causal understanding of a social concept: A methodological and statistical comparison of two methods," *Papers on Social Representations*, 2003, *12*, 2–1.

Griffiths, Thomas L. and Joshua B. Tenenbaum, "Structure and strength in causal induction," *Cognitive Psychology*, 2005, *51* (4), 334–384.

_ and _ , "Theory-based causal induction," *Psychological Review*, 2009, *116* (4), 661.

_ , Nick Chater, and Joshua Tenenbaum, *Bayesian Models of Cognition: Reverse Engineering the Mind*, MIT Press, 2024.

Hagmayer, York and Björn Meder, "Repeated Causal Decision Making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2013, *39* (1), 33–50.

_ and Michael R. Waldmann, "Inferences about unobserved causes in human contingency learning," *Quarterly Journal of Experimental Psychology*, 2007, *60* (3), 330–355.

_ and Philip M. Fernbach, "Causal Reasoning in Decision Making," in Michael R. Waldmann, ed., *The Oxford Handbook of Causal Reasoning*, Oxford University Press, 2017, pp. 543–562.

_ and Steven A Sloman, "Decision makers conceive of their choices as interventions.," *Journal of Experimental Psychology: General*, 2009, *138* (1), 22.

Halpern, Joseph Y. and Evan Piermont, "A representation theorem for causal decision making," in "Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning," Vol. 21 2024, pp. 410–419.

Hamilton, David L. and Robert K. Gifford, "Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments," *Journal of Experimental Social Psychology*, 1976, *12* (4), 392–407.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning through noticing: Theory and evidence from a field experiment," *Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic expectations and misguided learning," *Econometrica*, 2018, *86* (4), 1159–1214.

_ , _ , and _ , "Overconfidence and prejudice," *Review of Economic Studies*, 2025, p. rdaf048.

Holyoak, Keith J., Hee Seung Lee, and Hongjing Lu, "Analogical and category-based inference: a theoretical integration with Bayesian causal models," *Journal of Experimental Psychology: General*, 2010, *139* (4), 702.

Icard, Thomas F. and Noah D. Goodman, "A resource-rational approach to the causal frame problem," in "Proceedings of the Annual Meeting of the Cognitive Science Society," Vol. 37 2015.

Ichihashi, Shota and Delong Meng, "The design and interpretation of information," *Available at SSRN 3966003*, 2021.

Jain, Atulya, "Informing agents amidst biased narratives," Technical Report, Mimeo 2023.

Jehiel, Philippe, "Analogy-based expectation equilibrium," *Journal of Economic Theory*, 2005, *123* (2), 81–104.

Jenkins, H. M. and W. C. Ward, "Judgment of contingency between responses and outcomes," *Psychological Monographs: General and Applied*, 1965, *79* (1), 1–17. Whole No. 594.

Jern, Alan, Kai-Min K Chang, and Charles Kemp, "Belief polarization is not always irrational," *Psychological Review*, 2014, *121* (2), 206.

Jones, Edward E and Victor A Harris, "The attribution of attitudes," *Journal of Experimental Social Psychology*, 1967, *3* (1), 1–24.

Kahneman, Daniel and Amos Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, 1979, *47* (2), 263–291.

Keil, Frank C., "Folkscience: Coarse interpretations of a complex reality," *Trends in Cognitive Sciences*, 2003, *7* (8), 368–373.

Kemp, Charles, Joshua B. Tenenbaum, Sourabh Niyogi, and Thomas L. Griffiths, "A probabilistic model of theory formation," *Cognition*, 2010, *114* (2), 165–196.

_ , Noah D. Goodman, and Joshua B. Tenenbaum, "Learning to learn causal models," *Cognitive Science*, 2010, *34* (7), 1185–1243.

Kendall, Chad W. and Constantin Charles, "Causal narratives," *Unpublished*, 2022.

Khemlani, Sangeet S. and Daniel M. Oppenheimer, "When one model casts doubt on another: A levels-of-analysis approach to causal discounting.," *Psychological Bulletin*, 2011, *137* (2), 195.

Kinney, David and Tania Lombrozo, "Building compressed causal models of the world," *Cognitive Psychology*, 2024, *155*, 101682.

Koller, Daphne and Nir Friedman, *Probabilistic graphical models: principles and techniques*, MIT Press, 2009.

Kolvoort, Ivar R., Elizabeth L. Fisher, Robert van Rooij, Katrin Schulz, and Leendert van Maanen, "Probabilistic causal reasoning under time pressure," *PLoS ONE*, 2024, *19* (4), e0297011.

_ , Nina Temme, and Leendert van Maanen, "The Bayesian mutation sampler explains distributions of causal judgments," *Open Mind*, 2023, *7*, 318–349.

Körding, Konrad P., Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B. Tenenbaum, and Ladan Shams, "Causal inference in multisensory perception," *PLoS ONE*, 2007, *2* (9), e943.

Kőszegi, Botond and Matthew Rabin, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics*, 2006, *121* (4), 1133–1165.

Krynski, Tevye R. and Joshua B. Tenenbaum, "The role of causality in judgment under uncertainty.," *Journal of Experimental Psychology: General*, 2007, *136* (3), 430.

Kupor, Daniella and Kristin Laurin, "Probable cause: The influence of prior probabilities on forecasts and perceptions of magnitude," *Journal of Consumer Research*, 2020, *46* (5), 833–856.

Kushnir, Tamar, Alison Gopnik, Christopher G. Lucas, and Laura Schulz, "Inferring hidden causal structure," *Cognitive Science*, 2010, *34* (1), 148–160.

Lagnado, David A. and Steven A. Sloman, "Learning causal structure," in "Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society" Erlbaum Mahwah, NJ 2002, pp. 560–565.

_ and _ , "The advantage of timely intervention," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2004, *30* (4), 856–876.

Lagnado, David A and Steven A Sloman, "Time as a guide to cause.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2006, *32* (3), 451.

Lagnado, David A. and Tobias Gerstenberg, "Causation in legal and moral reasoning," in Michael R. Waldmann, ed., *The Oxford Handbook of Causal Reasoning*, Oxford: Oxford University Press, 2017, pp. 565–601.

_ , Michael R. Waldmann, York Hagmayer, and Steven A. Sloman, "Beyond Covariation: Cues to Causal Structure," in Alison Gopnik and Laura Schulz, eds., *Causal Learning: Psychology, Philosophy, and Computation*, Oxford University Press, 2007, pp. 154–172.

_ , Tobias Gerstenberg, and Ro'i Zultan, "Causal responsibility and counterfactuals," *Cognitive Science*, 2013, *37* (6), 1036–1073.

Langer, Ellen J., "The illusion of control," *Journal of Personality and Social Psychology*, 1975, *32* (2), 311.

Langer, Ellen J and Jane Roth, "Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task.," *Journal of Personality and Social Psychology*, 1975, *32* (6), 951.

Lejarraga, Tomás and Ralph Hertwig, "How experimental methods shaped views on human competence and rationality.," *Psychological Bulletin*, 2021, *147* (6), 535.

Levy, Gilat, Ronny Razin, and Alwyn Young, "Misspecified politics and the recurrence of populism," *American Economic Review*, 2022, *112* (3), 928–962.

Lombrozo, Tania and Nadya Vasilyeva, "Causal explanation," *Oxford Handbook of Causal Reasoning*, 2017, pp. 415–432.

Lu, Hongjing, Alan L. Yuille, Mimi Liljeholm, Patricia W. Cheng, and Keith J. Holyoak, "Bayesian generic priors for causal learning," *Psychological Review*, 2008, *115* (4), 955–984.

Lucas, Christopher G. and Thomas L. Griffiths, "Learning the form of causal relationships using hierarchical Bayesian models," *Cognitive Science*, 2010, *34* (1), 113–147.

MacKay, David J.C., *Information theory, inference and learning algorithms*, Cambridge University Press, 2003.

Marchant, Nicolás, Tadeg Quillien, and Sergio E. Chaigneau, "A context-dependent Bayesian account for causal-based categorization," *Cognitive Science*, 2023, *47* (1), e13240.

Matute, Helena, Fernando Blanco, Ion Yarritu, Marcos Díaz-Lago, Miguel A Vadillo, and Itxaso Barberia, "Illusions of causality: How they bias our everyday thinking and how they could be reduced," *Frontiers in Psychology*, 2015, *6*, 146427.

Mayrhofer, Ralf and Michael R. Waldmann, "Sufficiency and necessity assumptions in causal structure induction," *Cognitive Science*, 2015, *39* (4), 757–777.

_ and _ , "Sufficiency and necessity assumptions in causal structure induction," *Cognitive Science*, 2016, *40* (8), 2137–2150.

Meder, Björn and Ralf Mayrhofer, "Inference from explanations and causal knowledge," *Oxford Handbook of Causal Reasoning*, 2017, pp. 433–456.

Molavi, Pooya, Alireza Tahbaz-Salehi, and Andrea Vedolin, "Model complexity, expectations, and asset prices," *Review of Economic Studies*, 2024, *91* (4), 2462–2507.

Morris, Adam, Jonathan Phillips, Thomas Icard, Joshua Knobe, Tobias Gerstenberg, and Fiery Cushman, "Judgments of actual causation approximate the effectiveness of interventions," *Psy ArXiv) doi*, 2018, *10*.

Morris, Michael W. and Richard P. Larrick, "When one cause casts doubt on another: A normative analysis of discounting in causal attribution," *Psychological Review*, 1995, *102* (2), 331.

Musolff, Robin and Florian Zimmermann, "Model uncertainty," 2025. Working paper, version March 27, 2025.

Nam, Andrew, Christopher Hughes, Thomas Icard, and Tobias Gerstenberg, "Show and tell: Learning causal structures from observations and explanations," in "Proceedings of the Annual Meeting of the Cognitive Science Society," Vol. 45 2023.

Ng, David W., Jessica C. Lee, and Peter F. Lovibond, "The role of prior beliefs in causal illusions," *Cognition*, 2026, *266*, 106290.

Osborne, Martin J. and Ariel Rubinstein, "Games with procedurally rational players," *American Economic Review*, September 1998, *88* (4), 834–847.

Papakonstantinou, Trisevgeni, Kuan Iao Leong, and David Lagnado, "Humans generate auxiliary hypotheses to resolve conflicts in observational data," in "Proceedings of the Annual Meeting of the Cognitive Science Society," Vol. 46 2024.

Park, Juhwa and Steven A. Sloman, "Mechanistic beliefs determine adherence to the Markov property in causal reasoning," *Cognitive Psychology*, 2013, *67* (4), 186–216.

Pearl, Judea, *Causality*, Cambridge University Press, 2009.

_ , "The do-calculus revisited," *arXiv preprint arXiv:1210.4852*, 2012.

_ **and Dana Mackenzie**, *The book of why: the new science of cause and effect*, Basic Books, 2018.

Phelps, Edmund S., "Phillips curves, expectations of inflation and optimal unemployment over time," *Economica*, 1967, *34* (135), 254–281.

Pilgrim, Charlie, Adam Sanborn, Eugene Malthouse, and Thomas T. Hills, "Confirmation bias emerges from an approximation to Bayesian reasoning," *Cognition*, 2024, *245*, 105693.

Plach, Marcus, "Bayesian networks as models of human judgement under uncertainty: The role of causal assumptions in belief updating," *Kognitionswissenschaft*, 1999, *8* (1), 30–39.

Powell, Derek, Kara Weisman, and Ellen M. Markman, "Modeling and leveraging intuitive theories to improve vaccine attitudes," *Journal of Experimental Psychology: General*, 2023, *152* (5), 1379–1395.

Priniski, John, Ishaan Verma, and Fred Morstatter, "Pipeline for modeling causal beliefs from natural language," in Danushka Bollegala, Ruihong Huang, and Alan Ritter, eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2023, pp. 436–443.

Rehder, Bob, "Independence and dependence in human causal reasoning," *Cognitive Psychology*, 2014, *72*, 54–107.

_ , "Concepts as causal models: Categorization," in Michael R. Waldmann, ed., *The Oxford Handbook of Causal Reasoning*, Oxford University Press, 2017, chapter 20, pp. 347–376.

_ , "Concepts as causal models: Induction," in Michael R. Waldmann, ed., *The Oxford Handbook of Causal Reasoning*, Oxford University Press, 2017, chapter 21, pp. 377–414.

_ , "A rational process model of reasoning causally with continuous variables," *Cognition*, 2025, *263*, 106193.

_ **and Michael R. Waldmann**, "Failures of explaining away and screening off in described versus experienced causal learning scenarios," *Memory & Cognition*, 2017, *45* (2), 245–260.

_ **and** _ , "Failures of explaining away and screening off in described versus experienced causal learning scenarios," *Memory & Cognition*, 2017, *45* (2), 245–260.

_ **and Reid Hastie**, "Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity.," *Journal of Experimental Psychology: General*, 2001, *130* (3), 323.

_ **and Russell C. Burnett**, "Feature inference and the causal structure of categories," *Cognitive Psychology*, 2005, *50* (3), 264–314.

Rich, Alexander S. and Todd M. Gureckis, "The limits of learning: Exploration, generalization, and the development of learning traps," *Journal of Experimental Psychology: General*, 2018, *147* (11), 1553–1570.

Richens, Jonathan and Tom Everitt, "Robust agents learn causal world models," *arXiv preprint arXiv:2402.10877*, 2024.

Rose, David, Eric Sievers, and Shaun Nichols, "Cause and burn," *Cognition*, 2021, *207*, 104517.

Rothe, Anselm, Ben Deverett, Ralf Mayrhofer, and Charles Kemp, "Successful structure learning from observational data," *Cognition*, 2018, *179*, 266–297.

Rottman, Benjamin M., "The acquisition and use of causal structure knowledge," in M.R. Waldmann, ed., *The Oxford Handbook of Causal Reasoning*, Oxford University Press, 2017.

_ **and Reid Hastie**, "Reasoning about causal relationships: Inferences on causal networks," *Psychological Bulletin*, 2014, *140* (1), 109.

_ **and** _ , "Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away," *Cognitive Psychology*, 2016, *87*, 88–134.

_ **and Y. Zhang**, "Learning about causal relations that change over time: Primacy and recency over long timeframes in causal judgments and memory," *Cognitive Research: Principles and Implications*, 2025, *10* (1), 9.

_ **, Dedre Gentner, and Micah B. Goldwater**, "Causal systems categories: Differences in novice and expert categorization of causal phenomena," *Cognitive Science*, 2012, *36* (5), 919–932.

_ **, Woo-kyoung Ahn, and Christian C. Luhmann**, "When and how do people reason about unobserved causes?," in Phyllis McKay Illari, Federica Russo, and Jon Williamson, eds., *Causality in the Sciences*, Oxford: Oxford University Press, 2011, pp. 150–174.

Rozenblit, Leonid and Frank C. Keil, "The misunderstood limits of folk science: An illusion of explanatory depth," *Cognitive Science*, 2002, *26* (5), 521–562.

Samuelson, Larry and Jakub Steiner, "Robust latent representations," *Unpublished*, 2025.

Sargent, Thomas J., *The Conquest of American Inflation*, Princeton University Press, 1999.

Sato, Yoshiyuki, Taro Toyoizumi, and Kazuyuki Aihara, "Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli," *Neural Computation*, 2007, *19* (12), 3335–3355.

Schenone, Pablo, "Causality: A decision theoretic approach," *arXiv preprint arXiv:1812.07414*, 2024.

Schumacher, Heiner and Heidi Christina Thysen, "Equilibrium contracts and boundedly rational expectations," *Theoretical Economics*, 2022, *17* (1), 371–414.

Schwartzstein, Joshua, "Selective attention and learning," *Journal of the European Economic Association*, 2014, *12* (6), 1423–1452.

_ and Adi Sunderam, "Using models to persuade," *American Economic Review*, 2021, *111* (1), 276–323.

_ and _ , "Shared models in networks, organizations, and groups," *Unpublished*, 2022.

Shams, Ladan and Ulrik Beierholm, "Bayesian causal inference: A unifying neuroscience theory," *Neuroscience & Biobehavioral Reviews*, 2022, *137*, 104619.

Shanahan, Murray, "The Frame Problem," 2016. First published Mon Feb 23, 2004; substantive revision Mon Feb 8, 2016.

Shiller, Robert J, "Narrative economics," *American Economic Review*, 2017, *107* (4), 967–1004.

Siegal, Elitzur A. Bar-Asher, "Causation and causal relations: Insights from natural language," *Annual Review of Linguistics*, 2026, *12*, 125–145.

Sloman, Steven, *Causal models: How people think about the world and its alternatives*, Oxford University Press, 2005.

Sloman, Steven A. and York Hagmayer, "The Causal Psycho-Logic of Choice," *Trends in Cognitive Sciences*, 2006, *10* (9), 407–412.

_ , Aron K. Barbey, and Jared M. Hotaling, "A causal model theory of the meaning of "cause," "enable," and "prevent"," *Cognitive Science*, 2009, *33* (1), 21–50.

Spiegler, Ran, "Bayesian networks and boundedly rational expectations," *Quarterly Journal of Economics*, 2016, *131* (3), 1243–1290.

_ , "Behavioral implications of causal misperceptions," *Annual Review of Economics*, 2020, *12*, 81–106.

_ , "Can agents with causal misperceptions be systematically fooled?," *Journal of the European Economic Association*, 2020, *18* (2), 583–617.

_ , "On the behavioral consequences of reverse causality," *European Economic Review*, 2022, *149*, 104258.

_ , "Behavioral causal inference," *Review of Economic Studies*, 2025, p. rdaf050.

Spirtes, Peter, Clark N. Glymour, and Richard Scheines, *Causation, prediction, and search*, MIT Press, 2000.

Stephan, Simon, Neele Engelmann, and Michael R Waldmann, "The perceived dilution of causal strength," *Cognitive Psychology*, 2023, *140*, 101540.

Steyvers, Mark, Joshua B. Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum, "Inferring causal networks from observations and interventions," *Cognitive Science*, 2003, *27* (3), 453–489.

Sussman, Abigail B. and Daniel M. Oppenheimer, "A causal model theory of judgment," in "Proceedings of the Annual Meeting of the Cognitive Science Society," Vol. 33 2011.

_ and _ , "The effect of effects on effectiveness: A boon-bane asymmetry," *Cognition*, 2020, *199*, 104240.

Tatlidil, Semir, Steven A. Sloman, Semanti Basu, Tiffany Tran, Serena Saxena, Moon Hwan Kim, and Iris Bahar, "A comparison of methods to elicit causal structure," *Frontiers in Cognition*, 2025, *4*, 1544387.

Tenenbaum, Joshua and Thomas Griffiths, "Structure learning in human causal induction," *Advances in Neural Information Processing Systems*, 2000, *13*.

Trautmann, Stefan T. and Gijs Van De Kuilen, "Ambiguity Attitudes," in Gideon Keren and George Wu, eds., *The Wiley Blackwell Handbook of Judgment and Decision Making*, Vol. 2, Wiley, 2015, pp. 89–116.

Trommershäuser, Julia, Konrad Körding, and Michael S. Landy, *Sensory cue integration*, Oxford University Press, 2011.

Tversky, Amos and Daniel Kahneman, "Causal schemas in judgments under uncertainty," in "Progress in Social Psychology," Hillsdale: Erlbaum, 1980, pp. 49–72.

Verma, Thomas S. and Judea Pearl, "Equivalence and synthesis of causal models," in "Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence" Association for Uncertainty in Artificial Intelligence Mountain View, CA 1990, pp. 220–227.

Waldmann, Michael, *The Oxford Handbook of Causal Reasoning*, Oxford University Press, 2017.

Waldmann, Michael R. and Keith J. Holyoak, "Predictive and diagnostic learning within causal models: Asymmetries in cue competition," *Journal of Experimental Psychology: General*, 1992, *121* (2), 222.

_ , _ , and Angela Fratianne, "Causal models and the acquisition of category structure," *Journal of Experimental Psychology: General*, 1995, *124* (2), 181.

Wallsten, Thomas S. and David V. Budescu, "State of the art—encoding subjective probabilities: A psychological and psychometric review," *Management Science*, 1983, *29* (2), 151–173.

White, Peter A., "How well is causal structure inferred from cooccurrence information?," *European Journal of Cognitive Psychology*, 2006, *18* (3), 454–480.

Willett, C. L. and Benjamin M. Rottman, "The accuracy of causal learning over long timeframes: An ecological momentary experiment approach," *Cognitive Science*, 2021, *45* (7), e12985.

Yang, Jie, Soyeon Caren Han, and Josiah Poon, "A survey on extraction of causal relations from natural language text," *Knowledge and Information Systems*, 2022, *64* (5), 1161–1186.

Yarritu, Ion, Matute Helena, and A Vadillo Miguel, "Illusion of control. The role of personal involvement," *Experimental Psychology*, 2014, *61* (1), 38.

Yeung, Saiwing and Thomas L. Griffiths, "Identifying expectations about the strength of causal relationships," *Cognitive Psychology*, 2015, *76*, 1–29.

Zhang, Yiwen and Benjamin M. Rottman, "Causal learning with delays up to 21 hours," *Psychonomic Bulletin & Review*, 2024, *31* (1), 312–324.

Zhang, Yue and Gabriele Paolacci, "More correlations signal causation: The effect of correlational scope on perceived causality," *Journal of Consumer Research*, 2026, p. ucag005.