Cognition 277 (2026) 106684

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full length article

Manipulating prior causal beliefs affects the formation of apparent causal illusions☆

Simon Stephan *, Michael R. Waldmann

Department of Psychology, University of Göttingen, Gosslerstrasse 14, 37073 Göttingen, Germany

ARTICLE INFO

Dataset link: https://github.com/SimonStephan31/Priors_and_Causal-Illusions

Keywords:
Causal illusions
Outcome-density bias
Causal induction
Bayesian updating
Computational modeling
Mechanisms

ABSTRACT

Non-contingent learning data can produce an apparent "illusory" or "false" perception of causality ("causal illusion"), especially if a potential cause and effect frequently co-occur — a phenomenon known as the "outcome-density bias." In contrast to the prominent bias view, Bayesian models of causal induction explain these causal illusions as the result of a *rational* learning process in which observed data fail to fully override non-zero causal priors. Convincing evidence for this rational explanation requires an experimentally manipulated effect of causal priors — which has so far been lacking. We report four experiments ($N = 1860$) supporting the Bayesian interpretation. We successfully manipulated participants' causal priors either through visually conveyed causal mechanism information or through statistical base-rate information, and found that both manipulations influenced the degree to which participants reported a causal relation after having processed non-contingent learning data. The results were in line with Bayesian updating: lower causal priors yielded lower post-learning beliefs in a causal connection, effectively reducing the "causal illusion." The results strengthen computational models implementing a rational Bayesian view of causal induction.

1. Introduction

Causal induction is the process by which a reasoner learns whether one type of event causes another. A prominent view holds that causal learning to a large extent rests on the assessment of contingency between a potential cause and effect (see Le Pelley et al., 2017; Perales et al., 2017; Rottman, 2017, for overviews). Other factors, such as temporal contiguity, physical contact, or mechanism knowledge also influence causal learning beyond covariation (see Lagnado et al., 2007, for an overview). Consider a two-year-old child who receives a Toniebox for Christmas along with several other toys. Only one toy, a Tonie, makes the box play music when placed on top. The child systematically places each toy on the box: the first two produce nothing, even after repeated attempts; the third reliably activates it, establishing a positive contingency. When asked to make the box play music, the child promptly reaches for that toy.

This scenario illustrates that contingency-based causal induction often succeeds. Yet cases of apparent failure — so-called "causal illusions" — have also attracted sustained research attention (Allan & Jenkins, 1983; Buehner et al., 2003; Chow et al., 2019; Le Pelley et al., 2017; Musca et al., 2010; Shanks, 1995; Shanks et al., 1996; Vadillo et al., 2011; White, 2003). Participants who are shown zero-contingency learning data, in which there is no correlation between a putative cause and effect (i.e., $\Delta P = 0$), sometimes tend to report non-zero causal ratings — hence the label "causal illusion." This effect has also been studied in both healthy and clinical populations under the labels "illusion of control" or "depressive realism," and is likewise interpreted as a non-normative bias (e.g., Alloy & Abramson, 1979; Msetfi et al., 2005).

An example data set found to elicit causal illusions is shown in the left contingency table in Fig. 1. The effect (E) is always present when the cause (C) is present, $P(e^+ \mid c^+) = 1.0$, and also always present when it is absent, $P(e^+ \mid c^-) = 1.0$. There is no contingency between C and E ($\Delta P = P(e^+ \mid c^+) - P(e^+ \mid c^-) = 0$). Despite this absence of a positive covariation, participants tend to report causal judgments greater than zero, thus expressing an alleged illusory perception of causality. The effect tends to be strongest when $P(e^+ \mid c^-) = 1.0$ and to become weaker when $P(e^+ \mid c^-)$ decreases (cf. the middle and right contingency tables in Fig. 1): the degree to which a learner experiences the apparent causal

☆ All experimental materials, data, and R scripts (R version 4.5.1) are publicly available in a GitHub repository at https://github.com/SimonStephan31/Priors_and_Causal-Illusions. The studies' pre-registrations are available under: (Exp. 1) https://osf.io/ez4qm/overview, (Exp. 2) https://osf.io/r8xca/overview, (Exp. 3) https://osf.io/jdqgm/overview, (Exp. 4) https://osf.io/jc6x5/overview.

This work was supported by a Reinhart Koselleck project (WA 621/25-1), funded by the Deutsche Forschungsgemeinschaft (DFG).

* Corresponding author.
E-mail address: simon.stephan@psych.uni-goettingen.de (S. Stephan).

https://doi.org/10.1016/j.cognition.2026.106684
Received 10 February 2026; Received in revised form 4 August 2026; Accepted 6 August 2026
Available online 17 August 2026

	$N(e^+)$	$N(e^-)$
$N(c^+)$	4	0
$N(c^-)$	4	0

$P(e^+ \mid c^+) = 1.0$
$P(e^+ \mid c^-) = 1.0$
$\Delta P = 0$

	$N(e^+)$	$N(e^-)$
$N(c^+)$	2	2
$N(c^-)$	2	2

$P(e^+ \mid c^+) = 0.5$
$P(e^+ \mid c^-) = 0.5$
$\Delta P = 0$

	$N(e^+)$	$N(e^-)$
$N(c^+)$	0	4
$N(c^-)$	0	4

$P(e^+ \mid c^+) = 0$
$P(e^+ \mid c^-) = 0$
$\Delta P = 0$

Fig. 1. Examples of contingency data sets with a zero contingency ($\Delta P = 0$) between putative cause C and effect E. $^+$ signs denote a variable's presence and $^-$ signs its absence.

illusion is influenced by the outcome density $P(e^+ \mid c^-)$, which is why the effect has also been called "outcome-density bias."

What explains such an apparent causal illusion, and which factors other than outcome density may influence it? Accounts that characterize non-zero causal ratings after zero-contingency learning data as an irrational bias or an illusion often explain it in terms of differential weighting of trial types. A prominent weighting scheme (see Perales et al., 2017) is

$$N_i(e^+ \mid c^+) > N_i(e^- \mid c^+) \geq N_i(e^+ \mid c^-) > N_i(e^- \mid c^-)$$

(or $a > b \geq c > d$). According to this account, a causal illusion arises because learners assign disproportionate weight to confirming observations — especially trials in which C and E co-occur — while underweighting disconfirming ones.

While contingency-based approaches interpret observed covariations as reflecting objective statistical relations in the world, Bayesian causal learning theories focus on the rationality of judgments based on limited information. According to this approach, learners enter a task with prior assumptions about the world, which, with increasingly large samples, they update on the basis of what they observe. Adopting the Bayesian view on causal learning, Griffiths and Tenenbaum (2005) argued that the causal illusion effect admits a *rational* Bayesian explanation, formalized in computational models such as the causal support model (Griffiths & Tenenbaum, 2005) or the structure induction (SI) model (Meder et al., 2014; Stephan & Waldmann, 2018). On this view, non-zero causal judgments are not irrational illusions but arise normatively when a reasoner's assumption of a non-zero prior probability of a causal link is not fully overridden by the data. We therefore speak of "seeming" or "apparent" causal illusions in this paper, and adopt the neutral term "outcome-density *effect*" rather than "outcome-density *bias*" when speaking about the role of outcome density. The Bayesian analysis leads to the hypothesis that the magnitude of apparent causal illusions depends on the degree to which a reasoner believes that C causes E prior to seeing learning data.

A limitation of previous work on Bayesian models of causal learning as an explanation of apparent causal illusions is that causal structure priors have neither been directly measured nor successfully experimentally manipulated. A notable exception is a recent study by Ng et al. (2026), who measured participants' prior causal beliefs before presenting zero-contingency learning data. They found that participants held relatively strong prior beliefs in a causal connection, which may explain why post-learning causal judgments remained above zero. Crucially, however, their attempt to experimentally reduce these prior beliefs through different verbal scenario instructions was unsuccessful, limiting the extent to which the observed judgments could be attributed to prior beliefs rather than the learning data itself. In addition, participants' judgments were not compared to the quantitative predictions of a computational Bayesian causal learning model.

In a developmental study comparing adults and children, Griffiths et al. (2011) varied prior knowledge about causal relations in the context of a blicket detector paradigm, a task similar to our introductory Toniebox example in which participants learn which objects activate a detector (see Gopnik & Sobel, 2000). Causal priors were manipulated by introducing the participants to varying base rates of blickets and non-blickets (pencils containing or not containing "super lead" in the case of the adult participants). Adults' causal judgments were closely aligned with the quantitative predictions of a Bayesian model using varying prior probabilities over causal structures. In the task adapted for children, the results showed that 4-year-olds made causal judgments that qualitatively aligned with the Bayesian model predictions. While this work provided relevant supporting evidence for the influence of priors in causal learning, it differed from ours in key respects. Specifically, it employed a blocking paradigm and positive-contingency learning data (cf. Sobel et al., 2004). It did not address the formation of apparent causal illusions under zero contingencies or the role that outcome density plays in it.

It therefore remains an open empirical question whether differences in prior causal beliefs influence the magnitude of apparent causal illusions. We address this question and show empirically that they do. Moreover, we demonstrate that apparent causal illusions are practically eliminated in participants who hold a near-zero causal prior, and that it is strongest in those who strongly believe in a positive link prior to the data. Our findings corroborate the Bayesian learning account of causal induction and its explanation of causal illusions. We begin by summarizing the SI model framework as an example of a Bayesian causal learning model, and then derive its predictions, before presenting experiments in support of the Bayesian explanation of apparent causal illusions.

2. Illusory causation through the lens of Bayesian causal learning

Studies probing the apparent causal-illusion effect typically ask participants to rate the causal strength between a potential cause C and effect E after observing (zero-) contingency learning data. Participants' strength judgments are then evaluated against theories that model causal learning as the estimation of a point-valued causal strength directly computed from the observed data (Buehner et al., 2003; Cheng, 1997; Jenkins & Ward, 1965; Lober & Shanks, 2000; Shanks et al., 1996). Because these theories are treated as normative accounts, yet fail to predict non-zero causal ratings after zero-contingency data, participants' non-zero judgments are characterized as a genuine "bias", "causal illusion", or "false causal belief."

A different class of learning theories are Bayesian models of causal induction (Griffiths & Tenenbaum, 2005; Lu et al., 2008; Meder et al., 2014; Stephan & Waldmann, 2018), which go beyond point estimates obtained directly from the data. A core notion implemented in such models is that learners bring prior causal assumptions to the task, and that learning follows the principles of Bayesian updating (Griffiths & Tenenbaum, 2009). Bayesian learning models offer a rational explanation: post-learning causal judgments remain above zero because the observed evidence is insufficient to drive prior causal beliefs to zero. Moreover, higher outcome density in zero-contingency data provides weaker evidence *against* the existence of a causal link, yielding posterior judgments closer to the prior.

Fig. 2. Alternative causal structures (S_1 vs. S_0) explaining the observed contingency data. The target cause C and the effect E denote binary variables and the arrow parameters denote the link strengths. Node A represents background causes of E that are unobserved during learning. C is a cause of E under S_1 but not under S_0.

Bayesian frameworks differentiate between causal structure learning (Griffiths & Tenenbaum, 2005; Meder et al., 2014; Stephan & Waldmann, 2018) and causal strength learning (Lu et al., 2008). While structure learning models determine whether a causal link exists between two factors, strength learning models estimate the link's magnitude, contingent upon its existence. Although these two forms of inference can be dissociated through specific query formulations (Lu et al., 2008), researchers have argued that causal induction is often fundamentally a matter of structure learning (Griffiths & Tenenbaum, 2005, 2009; Sloman, 2005). The rationale is that human causal learning parallels the scientific method: people first ask whether a relationship exists before attempting to quantify it.

Although previous studies on causal illusions relied on causal strength queries, we nonetheless focus on Bayesian models of causal *structure* learning. We believe that the causal-illusion effect is better characterized as a structural phenomenon. This is supported by the definitions offered in the literature, which speak of the effect as follows: "[...] people develop the belief that there is a causal connection between two events that are actually unrelated" (Matute et al., 2015, p. 1), or "[...] a phenomenon that consists in believing that a causal relation exists between a potential cause, C, and an outcome, O, when they are causally unrelated but coincide frequently" (Blanco & Matute, 2019, p. 1). These are fundamentally claims about the *existence* of a link (structure) rather than its magnitude (strength).

Despite their conceptual difference, structural and strength intuitions are closely linked: believing that a link exists implies believing in a causal strength greater than zero, while believing that a link is absent implies a covariation of exactly zero. Consequently, participants in earlier studies have used strength scales as a proxy to express underlying structural beliefs (see also Griffiths & Tenenbaum, 2005).

Furthermore, a Bayesian model of strength learning would, on the broader conceptual level of Bayesian updating, still offer a rationalization similar to that of a structure-learning model: posterior strength ratings are not expected to collapse to zero if the data fail to provide sufficient evidence for a covariation (contingency) of zero. We will return to the question of how Bayesian models of causal strength learning would predict apparent causal illusions under zero-contingencies after our introduction of the Bayesian Structure Induction Model.

2.1. The structure induction model

We use the Structure Induction (SI) model (Meder et al., 2014; Stephan & Waldmann, 2018), originally inspired by the Causal Support model (Griffiths & Tenenbaum, 2005). Both models make identical assumptions: the SI model yields the posterior probability of a causal link, whereas the causal support model computes the log-likelihood ratio between structures with and without such a link, which can be converted to a posterior probability via a logistic transformation (cf. Griffiths & Tenenbaum, 2009).

2.1.1. Bayesian updating of structural hypotheses

According to the SI model, a causal learner uses empirical data as evidence to update prior probabilities over alternative causal hypotheses about the data. These hypotheses are formally represented as causal Bayes nets (Glymour, 2001; Pearl, 2000; Spirtes et al., 1993; Spohn, 2001), shown in Fig. 2. Under structure S_1, the causal arrow between C and E represents that C is a cause of E, in addition to unobserved independent background causes A. Some of the observed co-occurrences of C and E are thus explained by the causal influence of C; others are coincidences in which E is caused by an unobserved background cause A. Under the competing structure S_0, no causal link between C and E is assumed; E is influenced solely by A, and all observed co-occurrences are treated as coincidental. Learning updates prior structure probabilities in accordance with Bayes' rule:

$$P(S_i \mid D) = P(S_i)\frac{P(D \mid S_i)}{P(D)}, \tag{1}$$

where $P(S_i \mid D)$ denotes the posterior probability of structure S_i, $P(S_i)$ its prior probability, $P(D \mid S_i)$ the likelihood of the observed data under S_i, and $P(D)$ the marginal probability of the data D. The full formalization of the SI model can be found in Meder et al. (2014). Crucially, Eq. (1) entails that the posterior probability of a causal link between C and E is proportional to the product of the likelihood of the data under S_1 and its prior probability:

$$P(S_i \mid D) \propto P(S_i) \cdot P(D \mid S_i). \tag{2}$$

The default assumption of the SI model is that a learner approaches the task with an uninformative structure prior — that is, regarding S_0 and S_1 as equally plausible: $P(S_1) = P(S_0)$. Because S_1 and S_0 are mutually exclusive and exhaustive, $P(S_1) = 1 - P(S_0)$; a uniform prior thus implies $P(S_1) = P(S_0) = 0.5$.

Assuming that an uninformative structure prior captures participants' assumptions in previous studies, the SI model predicts alleged causal illusions. It also identifies the circumstances under which an influence of outcome density should be particularly strong or particularly weak. This is illustrated in Fig. 3. The graphs show how the posterior belief in S_1 ($P[S_1 \mid D]$, y-axis) changes with the prior belief ($P[S_1]$, x-axis) and the observed data D, for a range of positive and zero-contingency sets. Assuming an uninformative prior of 0.5 (midpoint on the x-axis), posterior beliefs in S_1 are clearly greater than zero and decrease as $P(e^+ \mid c^-)$ decreases. The visualizations show that this holds for S_1 priors other than 0.5 as well, although the influence of $P(e^+ \mid c^-)$ is attenuated for extreme priors.

Another relevant factor, not depicted in the graphs, is sample size: given a positive contingency, the posterior of S_1 increases more quickly with more data; under zero contingency, larger samples drive beliefs in S_1 downward (see Griffiths & Tenenbaum, 2005, for an empirical test). This means that any non-zero prior (unless it is 1.0 and thus immune to evidence) should eventually approach zero given sufficient exposure to a zero contingency.

The plausibility of an uninformative structure prior depends on the knowledge a learner brings to the task (cf. Griffiths & Tenenbaum, 2009). In typical cover stories — usually describing clinical trials testing newly developed drugs — S_1 priors close to 0.5 appear plausible. In contexts where domain-specific knowledge is available (cf. Griffiths & Tenenbaum, 2009), more informative structure priors are more appropriate. Fig. 3A shows continuous prior-posterior curves; Fig. 3B shows the change from prior to posterior probability for five highlighted structure prior values.

Any prior belief in S_1 greater than zero yields posterior beliefs greater than zero, with stronger posterior beliefs for stronger S_1 priors. Importantly, this is a normative pattern according to the model, not a

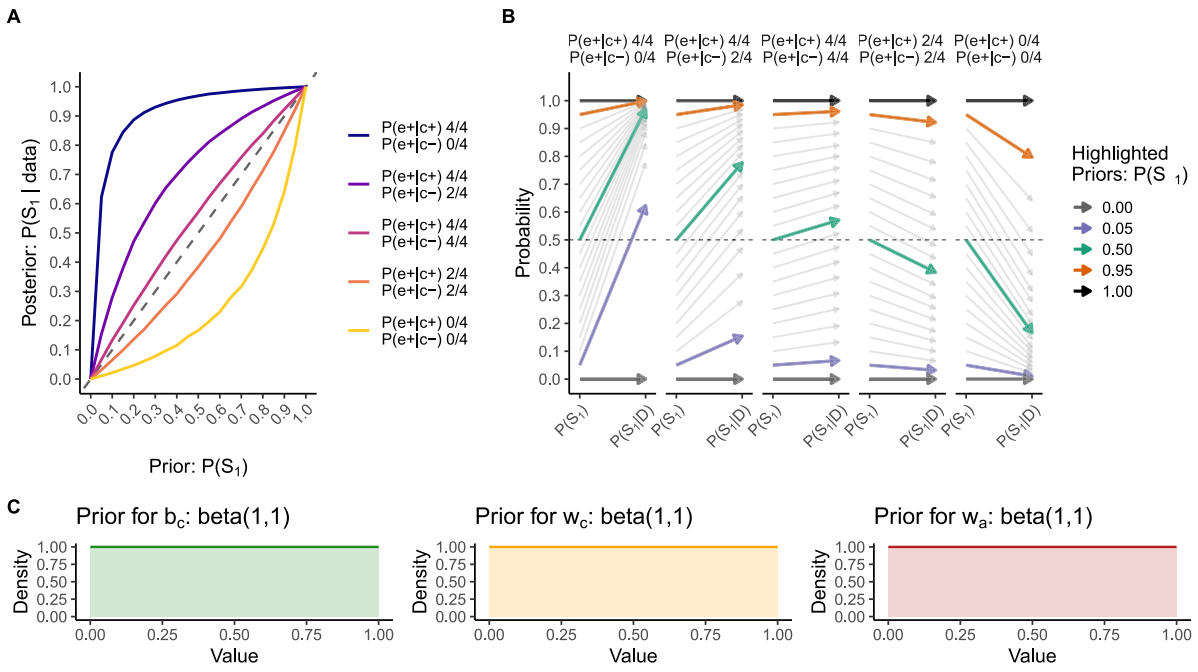

Fig. 3. SI model predictions. A: Prior-posterior curves showing the relation between the prior and posterior probability of S_1 for two positive and three zero contingencies. B: Prior-posterior changes under these contingencies for five different highlighted S_1 priors. C: Flat Beta(1, 1) priors over structure S_1's parameters. Under structure S_0, the only difference is that w_c is set to 0.

"causal illusion." As shown by the prior-posterior curves in Fig. 3A and the discrete prior-posterior differences in Fig. 3B, the posterior belief in S_1 also tends to be higher the weaker the evidence against S_1. For example, a learner observing $P(e^+|c^+) = P(e^+|c^-) = \frac{2}{4}$ holds a higher posterior belief in S_1 than one observing $P(e^+|c^+) = P(e^+|c^-) = \frac{0}{4}$, a data set providing stronger evidence against S_1. This is the typical pattern of the outcome-density effect.

A further prediction of the SI model is that the degree to which the *same* contingency data impacts a learner's posterior belief in S_1 depends on the magnitude of their prior belief. Fig. 3B shows that the magnitude of the difference between the prior and posterior probability of S_1, given a specific contingency data set, changes depending on the prior probability of S_1. The impact of the data is predicted to be strongest for intermediate structure priors around 0.5, and disappears as priors approach the extremes. The largest impact of outcome density on posterior causal beliefs is therefore expected in learners who are relatively uncertain about the underlying causal structure (S_0 vs. S_1). A seemingly counterintuitive prediction visible in Fig. 3B is that the change in belief in S_1 from prior to posterior is smaller for high S_1 prior values (e.g., $P(S_1) = 0.95$), even though a learner starting at 0.95 appears to have more "room to move" toward zero in light of disconfirming evidence (e.g., $P(e^+|c^+) = P(e^+|c^-) = \frac{0}{4}$).[1] This pattern is, however, a direct consequence of the multiplicative nature of Bayesian updating. Posterior probabilities are obtained by combining the likelihood of the data under each competing hypothesis with its prior probability and normalizing them across hypotheses. As a result, the impact of a given dataset on the posterior probability of S_1 depends not only on how well it is explained by each causal structure, but also on the prior odds with which inference begins. A learner with a prior of 0.95 enters the task with odds of 19 : 1 in favor of S_1. Even strongly disconfirming data must first overcome this large initial advantage before the posterior moves substantially. The same data applied to a prior of 0.50 produce a much larger absolute shift, because the prior offers no such anchor. In short, a strong prior belief does not create much room to move when disconfirming evidence is encountered — it creates resistance to movement. How strong this resistance is also depends on the specific parameter priors (see below).

Finally, Figs. 3A and B also show that, for the contingency data sets shown in the graphs, a near-zero posterior belief in S_1 — and thus the disappearance of a "causal illusion" — can only arise if a reasoner starts with a near-zero causal prior. The model likewise predicts that the influence of $P(e^+|c^-)$ disappears for extreme structure priors.

2.1.2. The role of parameter priors

Besides uncertainty about causal structure, the SI model also incorporates uncertainty about the structures' parameters b_c, w_c, and w_a. The parameters w_c and w_a represent the strengths of target cause C and alternative cause A, respectively; b_c represents the base rate probability of C. Our main focus is not on identifying which strength priors best capture learners' performance in a specific task (see Lu et al., 2008; Meder et al., 2014, for such analyses), but because the predictions of the SI model depend on both structure and strength priors, it is useful to illustrate their joint influence on the posterior probability of S_1. The chosen parameter priors determine the model's sensitivity to evidence, independent of the structure prior. For example, a learner may have specific beliefs about the strength of a link, if it existed, while simultaneously doubting that it does. Consider a person who tries to find the correct light switch in a concert hall: they may doubt they chose correctly, while believing that the correct switch, if found, would work reliably.

In its standard form, the SI model (Meder et al., 2014) uses independent flat Beta(1, 1) distributions as parameter priors (Fig. 3C). Conceptually, flat priors over w_c and w_a represent a learner who considers all possible causal strengths of C and A as equally plausible *a priori*; a flat prior over b_c represents equal plausibility for each possible base rate of C.

Although the default model using flat priors predicted participants' behavior well in previous experiments (Griffiths & Tenenbaum, 2005; Meder et al., 2014; Stephan & Waldmann, 2018) (but see also Lu et al., 2008, who obtained higher model fits using a "strong and sparse"

[1] We thank an anonymous reviewer for pointing out that this seemingly counterintuitive characteristic of Bayesian models should be explained.

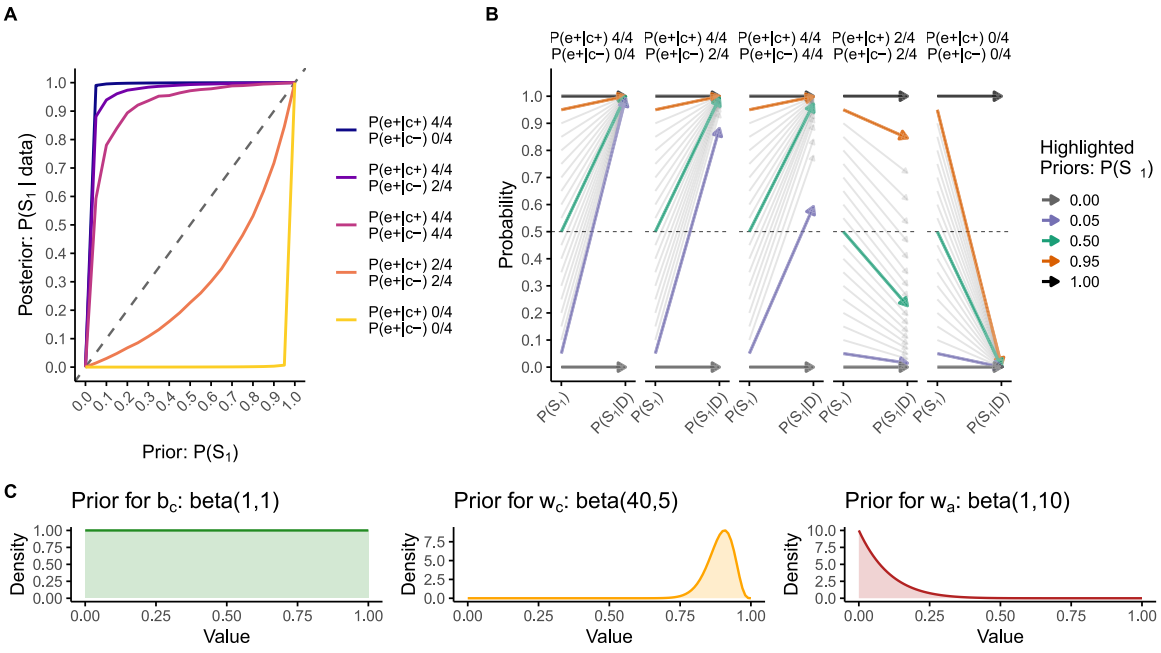

Fig. 4. SI model predictions with an informative w_c prior. A: Prior-posterior curves showing the relation between the prior and posterior probability of S_1 for two positive and three zero contingencies. B: Prior-posterior changes under these contingencies for five different highlighted S_1 priors. C: Flat Beta$(1, 1)$ prior over b_e, a Beta$(40, 5)$ prior over w_c, and a Beta$(1, 10)$ prior over w_a. Under structure S_0, the only difference is that w_c is set to 0.

prior), more informative parameter priors can model contexts in which a learner brings specific causal strength assumptions to the task.

Fig. 4 shows what happens under a set of strength priors that maximally distinguishes between S_1 and S_0: a prior over w_c biased towards high values, and a prior over w_a biased towards low values.[2] One domain where such assumptions may be appropriate is artifacts or devices: a learner may not know how a novel device works but may still assume that it can be activated reliably by the correct intervention (high w_c), and that it stays inactive otherwise (low w_a). We will consider this possibility because our studies use unfamiliar devices as a cover story.

Compared to the flat-prior case (Fig. 3), priors that maximally distinguish between S_1 and S_0 entail increased sensitivity to confirming and disconfirming evidence. Observing a zero contingency — especially one with a low outcome-density — provides strong evidence against a high strength of C. This leads to a pronounced drop in belief in a causal link. By contrast, observing a positive contingency produces a pronounced increase in belief in S_1. Comparing the predictions that this version of the model makes for zero-contingencies with the ones made by the model using flat parameter priors, it can be seen that the effect of outcome density on alleged causal illusions also depends on the prior assumptions about the causal parameters.

A noteworthy scenario (cf. Figs. 3, 4), is the ceiling condition, $P(e^+|c^+) = P(e^+|c^-) = \frac{4}{4}$. As panels A and B show, the SI model predicts an increase from prior to posterior unless S_1 priors take on extreme values. This happens because ceiling data leave the causal strength of C underdetermined (cf. Cheng, 1997): the observed effects could be generated entirely by background causes, A, or C could additionally contribute but be masked by A's influence. Under S_0, by contrast, w_c is fixed at 0, meaning the only way S_0 could generate ceiling data is through an alternative cause with deterministic strength $w_a = 1.0$. Because Bayesian model averaging integrates over parameter uncertainty, S_1 retains a broader range of parameter values consistent with the data than S_0 does. In short, ceiling data are ambiguous: they are consistent with S_1 across a wider parameter range than with S_0, so in the SI model Bayesian averaging sometimes favors S_1 even under zero contingency. Figs. 3 and 4 also show that the degree to which S_1 is favored under ceiling conditions depends on the prior parameter assumptions. Believing that w_c is high and that w_a is low boosts the effect. Conversely, under a set of parameter priors in which the w_c prior favors low values and the w_a prior favors high values, the posterior advantage of S_1 under ceiling conditions would practically disappear.

A strength prior favoring high w_c and low w_a values results in high sensitivity to the data (or outcome density), because such priors maximally distinguish between S_1 and S_0. The opposite holds when a learner believes that C, should it be linked to E at all, has very low causal strength: assuming a low w_c makes S_1 and S_0 very similar, and the likelihood of the data, $P(D|S_i)$, becomes nearly identical for both structures. As a specific example, consider a learner whose prior over w_c is modeled with a Beta$(1, 10)$ distribution (favoring small w_c values), and whose prior over w_a is modeled with Beta$(80, 80)$ (yielding a w_a peak around 0.5). These distributions are shown in Fig. 5C. Fig. 5A shows that because of the low w_c prior distribution the prior-posterior curves are much closer together and far less responsive to the data.

A final noteworthy pattern in Fig. 5 is that there can be situations in which the learning data provide no evidence for or against S_1, even when the structure priors are not extreme. As panels A and B show, this occurs when the observed data perfectly match the prior parameter expectations. In our example where $w_a \approx 0.5$ is expected, the data set $P(e^+|c^+) = P(e^+|c^-) = \frac{2}{4}$ instantiates this case. The corresponding prior-posterior curve falls directly on the diagonal: the data leave the learner unable to distinguish between S_0 and S_1 regardless of their structure prior.

As mentioned above, Bayesian models of causal strength learning also explain causal illusions under zero-contingency learning data as a result of (rational) updating of non-zero priors. Our goal is not to compare different classes of Bayesian models of causal learning, nor to settle whether learners care primarily about causal strength or structure. The point is that any Bayesian model of causal induction, whether it is concerned with strength or structure, offers the same kind of rational explanation for apparent "causal illusions" under zero-contingency data: assuming a learner holds non-zero causal (strength or

[2] We focus here on differences in w_c and w_a, keeping the flat prior distribution over b_e.

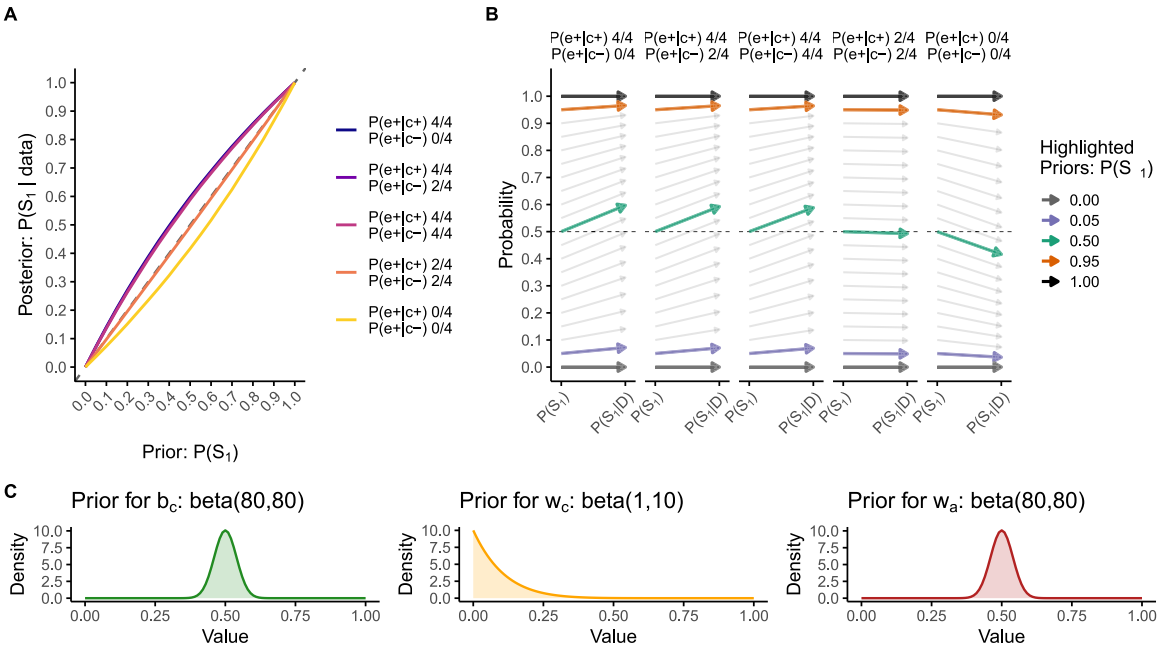

Fig. 5. SI model predictions with informative b_c, w_c, and w_a priors. A: Prior-posterior curves showing the relation between the prior and posterior probability of S_1 for two positive and three zero contingencies. B: Prior-posterior changes under these contingencies for five different highlighted S_1 priors. C: Informative beta priors over structure S_1's parameters b_c, w_c, and w_a. Under structure S_0, the only difference is that w_c is set to 0.

structure) priors, their posterior beliefs will remain greater than zero unless the data provide sufficiently strong evidence against this estimate.[3] Our illustration of the role of parameter priors in the SI model helps illustrate how a Bayesian strength learning model would predict apparent "causal illusions." Under one version of a Bayesian strength learning model, reasoners may use the learning data to update a prior distribution over the parameter w_c within structure S_1 (for a specific example of a model that considers only S_1 to model causal strength learning, see Lu et al., 2008). Such a model may use the mean of w_c's posterior distribution as the predicted value of a reasoner's causal strength judgment. Zero-contingency learning data will shift probability mass in the prior distribution over w_c toward zero. However, unless the prior distribution already strongly favors values close to zero, the posterior mean will remain greater than zero. In addition to the amount of data a learner observes (i.e., sample size), this posterior mean will also be influenced by outcome density, in a manner similar to the posterior probability of S_1 under the SI model: it will move closer to zero as outcome density decreases.

An alternative Bayesian way of modeling causal strength judgments, which would keep uncertainty about the underlying causal structure, would be to use Bayesian model averaging over S_0 and S_1 (see Meder et al., 2014, who modeled learners' diagnostic inferences in this way). Under this approach, one would compute the posterior mean of w_c separately for each structure (although note that w_c is fixed at 0 under S_0) and weigh each mean by the corresponding structure's posterior probability, $P(S_i|D)$. The final Bayesian posterior strength estimate would then correspond to the sum of these weighted posterior strength estimates. Such a model would also predict non-zero posterior (strength) judgments in light of (limited) zero-contingency data (unless the priors already favor zero).

3. Methods of manipulating priors

The analyses above illustrate that, regardless of which scenario is most plausible in a given context, the SI model predicts that a learner's apparent susceptibility to forming a "causal illusion" is, besides outcome density, influenced by their prior causal belief in S_1. We now turn to the question of how causal structure priors may be effectively manipulated experimentally.

3.1. Verbal cover stories

Typical causal learning experiments rely on *verbally instructed* (fictitious) cover stories, often about medical scenarios (see, e.g., Meder et al., 2014; Ng et al., 2026; Stephan & Waldmann, 2018). One possible approach to manipulate structure priors is therefore to vary the domain of the cover story. This was recently attempted by Ng et al. (2026), but their "attempt to reduce the positive prior belief by scenario instructions was unsuccessful" (p. 1). Beyond the difficulty of finding cover stories that reliably induce different priors, this approach offers limited experimental control: different cover stories may differ along dimensions other than the intended one.[4]

3.2. Statistical information

An alternative approach is the provision of explicit statistical information (see, e.g., Griffiths et al., 2011). For example, participants might be given summary statistics or "historical" data regarding the relationship between cause and effect. In a medical scenario, participants could be told how often a given substance causes a reaction in similar populations, based on prior trials. This method aims to establish a specific prior probability for S_1 versus S_0 through direct description rather than indirect association.

Although this approach offers high experimental control, it faces both empirical and theoretical challenges. First, explicit statistical information may lack the intuitive force of thematic background or mechanistic knowledge. Research on the base-rate fallacy (Kahneman & Tversky, 1973) suggests that learners frequently underweight descriptive priors (but see Koehler, 1996) once they encounter novel

[3] We thank an anonymous reviewer for the suggestion to say more about how Bayesian causal strength models would explain apparent causal illusions.

[4] Highly parallelized cover stories maximize control but may elicit similar priors, whereas highly different cover stories risk introducing confounds.

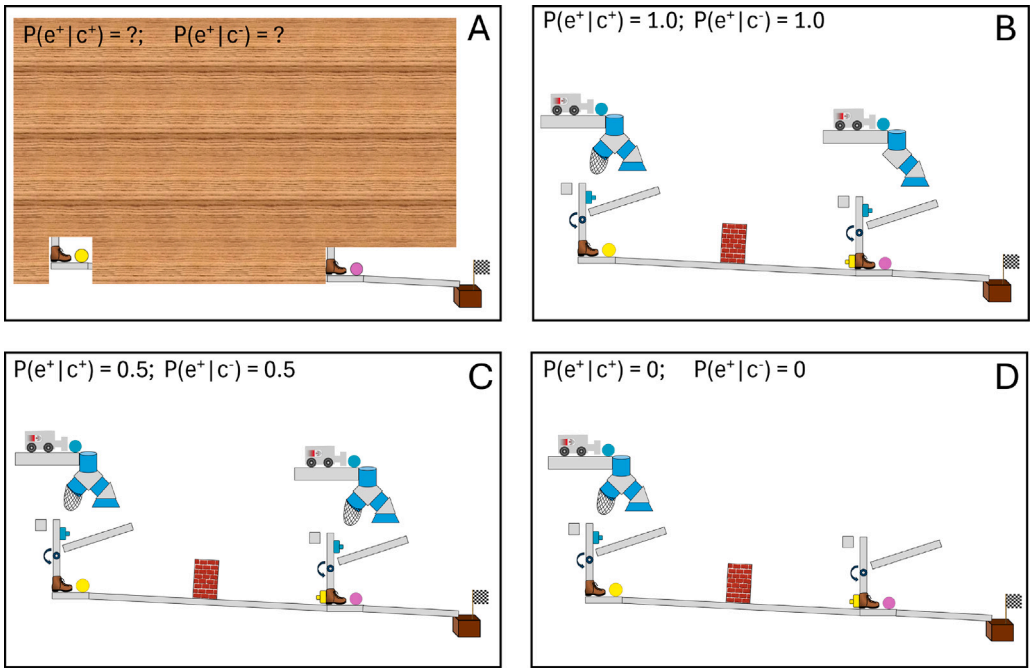

Fig. 6. Experimental stimuli used in Experiment 1. A: illustration of the device with its mechanism covered. This picture was shown to participants in the opaque mechanism condition. B–D: illustration of the device's mechanism in the different contingency conditions. The included probability information shows the contingencies the respective mechanisms plausibly generate. This information was not shown to participants.

information, such as eyewitness testimony (e.g., Bar-Hillel, 1980). Second, presenting priors using abstract probabilities may feel artificial, as real-world causal induction rarely begins with a pre-computed statistical summary (though such cases certainly exist). A related concern is that abstract statistical summaries may not translate into stable prior beliefs that are robustly maintained when participants encounter conflicting trial-by-trial evidence.

3.3. Visual mechanism information

We sought to manipulate causal priors by varying *visual causal mechanism* information, capitalizing on reasoners' ability to interpret visually conveyed causal mechanisms when these are presented in a tractable form. Our notion of a tractable visual mechanism representation is grounded in the New Mechanism view (Cartwright, 2019; Cartwright et al., 2020; Glennan, 2017; Illari & Williamson, 2012; Machamer et al., 2000), which characterizes a mechanism as follows: "A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon" (Glennan, 2017, p. 17). Entities involved in causal mechanisms are spatially arranged and possess stable capacities, dispositions, or powers; they can act and be acted upon, and are thereby involved in the production of change (Glennan, 2017). Fig. 6 shows our visual stimuli. Although the displays depict static devices, the processes they can give rise to are easily simulated based on knowledge about the entities, their capacities, and their configuration (see also Allen et al., 2020; Battaglia et al., 2013; Gerstenberg et al., 2021; Ullman et al., 2017, for accounts of "mental physical simulations").

Taking the New Mechanism view as a guide to how lay reasoners make sense of mechanisms, we constructed depictions of causal mechanisms at a resolution that renders the causally relevant *entities*, their *capacieties*, and relevant aspects of their *spatial arrangement* sufficiently salient to support a prior favoring S_0 — that is, ruling out a causal relation between cause and effect. Mechanism manipulations have previously also been used in studies testing how expectations about cause–effect contiguity influence causal learning (e.g., Buehner & May, 2002; Buehner & McGregor, 2006), showing that a seemingly stable bias for causal immediacy is moderated by underlying mechanism assumptions (see also Gong et al., 2026).

Our experimental scenario involves an unfamiliar kind of mechanical device allegedly used by psychologists to study causal learning. Fig. 6A depicts the device's external appearance with its mechanism hidden ("black box"/opaque condition); Figs. 6B–D show the mechanisms of different device versions presented in the transparent conditions.

The verbal scenario description explicitly mentions two device types: a causal one (instantiating S_1), in which rolling of the yellow ball can cause rolling of the pink ball, and a non-causal one (instantiating S_0), in which it cannot. The instructions also pointed out that the pink ball may roll even when the yellow ball does not, due to an alternative launching mechanism inside the device.

The mechanism illustrations depict a vehicle pushing a light-blue ball into a forking tube. The ball follows one of two paths: a left branch leading to a net or a right branch descending a slope that triggers a lever launching the yellow ball — a component common to all devices, intended to convey a 50% probability of the yellow ball rolling (C). Once launched, the yellow ball travels toward the pink ball and, absent obstruction, has the capacity to trigger a second lever that causes it to roll. However, a wall is positioned along this path, blocking the yellow ball from reaching the pink ball. The pink ball may nonetheless roll via a second triggering component whose design varies across device versions to convey different baseline probabilities of the effect, $P(e^+|c^-)$: 1.0 (Fig. 6B), 0.5 (Fig. 6C), and 0 (Fig. 6D). No additional information about how the mechanisms work was provided beyond these illustrations showing the static devices.

The mechanism illustrations in Fig. 6 are intended to suggest that a causal link between the yellow and pink ball is unlikely,[5] which (1) suggests a strong prior for S_0 and (2) renders subsequent zero-contingency learning data plausible. If this impression is indeed conveyed, and if structure priors influence the outcome-density effect, this effect should be diminished in participants who view these transparent devices.

[5] It remains possible in principle — for example, if the boot strikes the yellow ball in a way that it flies over the wall.

When participants cannot see how the device works, they start with more uncertain assumptions about its causal structure (opaque condition). Because of this uncertainty, after they observe evidence, they tend to report a stronger belief in a causal relationship than people who already understand the mechanism. Similarly, participants who are shown mechanism illustrations suggesting the presence of a link (i.e., high S_1 priors) should show the highest causal ratings.

Choosing appropriate parameter priors for generating model predictions depends on how participants engage with the mechanism illustrations — specifically, whether they focus on causal structure rather than component strength. In this case, behavior may be best captured by the default flat parameter priors, or by a set of domain-specific default priors. By contrast, if participants do attend to parameter information, a model using a low w_c prior and w_a prior values adapted to the specific mechanism components may be more appropriate.

4. Overview of experiments

We conducted four experiments to test the hypothesis that the outcome-density effect is a product of rational Bayesian updating and is therefore modulated by prior causal beliefs. In all experiments, we used the novel "ball-device" paradigm to manipulate the priors by showing participants the internal mechanical workings of the device. All experimental materials and data analysis files are available in the repository (https://github.com/SimonStephan31/Priors_and_Causal-Illusions). All studies were pre-registered on the OSF platform.

Experiment 1 tested the core prediction: if the causal-illusion effect is driven by non-zero priors, visual mechanism information suggesting the absence of a causal link (S_0) should mitigate the effect. We compared this "transparent" low-prior condition against an "opaque" condition suggesting an uninformative prior.

Experiment 2 aimed to rule out the possibility that visual mechanism information simply causes learners to ignore subsequent empirical data. We introduced a "high-prior" transparent condition (where the mechanism appeared fully functional) and included a positive-contingency condition to observe how participants update their beliefs when data and priors are incongruent. A further change was that in the transparent conditions we revealed the mechanism only once, prior to the test questions and the learning data.

Experiment 3 investigated the generalizability of our findings to causal strength test queries. While Experiments 1 and 2 used causal structure queries (i.e., probing participants' belief in the existence of a link), Experiment 3, similar to previous studies on causal illusions, employed a causal strength query to determine whether the influence of mechanism-based priors persists when participants are asked to quantify the magnitude of a causal influence.

Experiment 4 compared different techniques to manipulate priors. We contrasted our visual-mechanistic manipulation with a statistical-descriptive manipulation, providing base-rate information about the frequency of causal versus non-causal devices (cf. Griffiths et al., 2011). This allowed us to test whether the source of the prior — visually conveyed mechanism intuition versus numeric statistical information — alters subsequent Bayesian integration of contingency data. The comparison also allowed us to assess how much participants' judgments were driven by structural information (S_0 vs. S_1) versus information about causal strength parameters (e.g., w_A, b_C; Fig. 2).

A supplementary study (described in detail in the repository) provided an even more direct test of whether participants' judgments are primarily driven by the structural information conveyed by our mechanism illustrations or by the additional parameter information they contain. The supplementary study compared two conditions: one in which participants saw the full mechanism illustration (as in the previous experiments), and one in which they saw only the part of the illustration depicting the path between yellow and pink (with or without the wall, depending on condition) — the component most directly relevant to the structural question of whether a causal link between C and E exists. The results were very similar across the two conditions, suggesting that participants primarily focused on the structural information conveyed by the mechanism illustrations directly linking cause and effect, and that the additional parameter information about other components of the mechanism was considered only by a subset of participants.

5. Experiment 1

Experiment 1 provides an initial test of the hypothesis that apparent causal illusions are influenced by structural causal priors. We predicted that participants who are uncertain about the existence of a causal relation should provide higher causal ratings after zero contingency learning data (i.e., more pronounced apparent "causal illusions") than participants who regard the existence of a causal relation as *a priori* unlikely.

5.1. Methods

5.1.1. Participants

Four hundred and ninety-two participants were recruited via www.prolific.com. Participants had to be at least 18 years old, be native English speakers, have an approval rate of at least 90%, participate via a PC or laptop, not have taken part in pilot studies, and have at least some formal educational qualification (i.e., participants reporting "no formal qualifications" were excluded).[6] Two participants were excluded because they did not confirm participation via a PC or laptop. An additional 130 participants (26.5%) were excluded because they did not meet a prespecified learning-phase performance criterion (see Procedure and the pre-registration). The final analysis sample consisted of $N = 360$ participants (187 male, 168 female, 3 non-binary, 2 preferred not to say; $M_{age} = 42$ years, range 18–78 years). Data collection was terminated once all conditions included $n = 60$ participants after exclusions. The sample size was based on the results of a pilot study ($N = 300$), summarized in the repository, and an *a priori* power analysis detailed in the pre-registration.

5.1.2. Design, materials, and procedure

The study employed a 2 (device mechanism: opaque [suggesting neither S_0 nor S_1] vs. transparent [suggesting S_0]; between-subjects) × 3 (contingency: (4/4)(4/4) vs. (2/4)(2/4) vs. (0/4)(0/4); between-subjects) × 2 (causal rating type: prior vs. posterior; within-subjects) mixed design.

Participants read an instruction about an unfamiliar device used by psychologists who studied causal learning. They were told that the device generates movement of a yellow and a pink ball, and were shown a picture of the device with its mechanism covered (Fig. 6A). They also learned that in one device type, rolling of the yellow ball can cause the pink ball to roll (instantiating S_1), whereas in another type it cannot (instantiating S_0).

Participants were then informed that their task was to determine whether rolling of the yellow ball can cause rolling of the pink ball in the specific device shown at the top of the screen (cf. Fig. 6A). In the transparent mechanism conditions, participants were additionally shown an image revealing the device's mechanism, displayed at the bottom of the screen. Depending on the contingency condition, this was one of the stimuli in Fig. 6B–D, intended to influence participants' causal priors.

On a subsequent screen, participants answered an initial causal query assessing their structure prior. Depending on the device mechanism condition, the screen showed either the opaque device or the mechanism illustration (cf. Fig. 7). The query read: "How confident

[6] This criterion was applied because we wanted to ensure that participants are capable of reading and understanding the instructions.

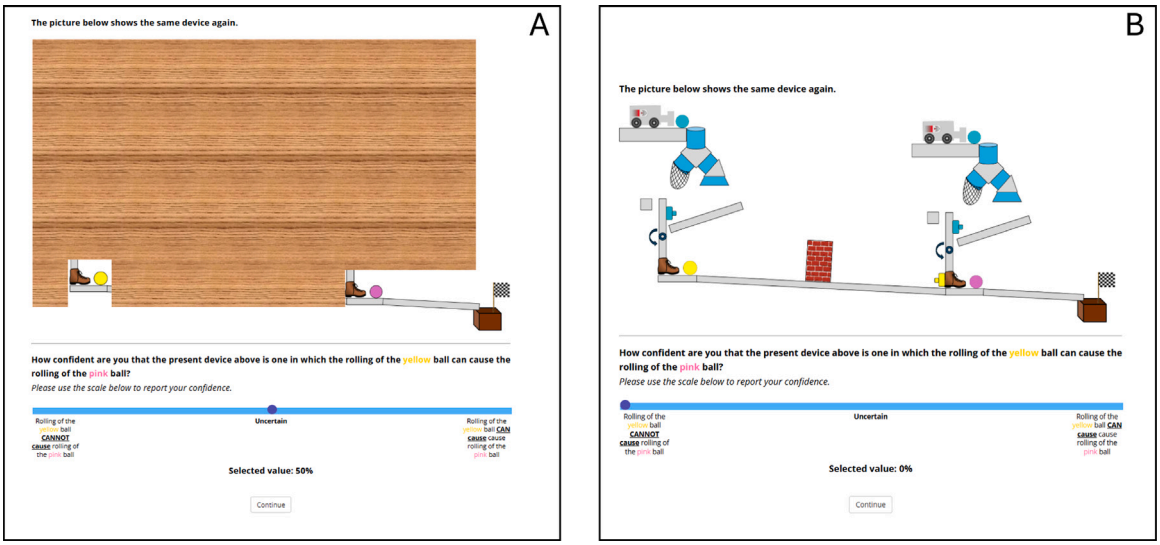

Fig. 7. Illustration of the rating screens shown in Experiment 1. A: example screen from the opaque mechanism conditions. B: example screen from the transparent mechanism condition where $P(e|c) = P(e|\neg c) = 0.5$.

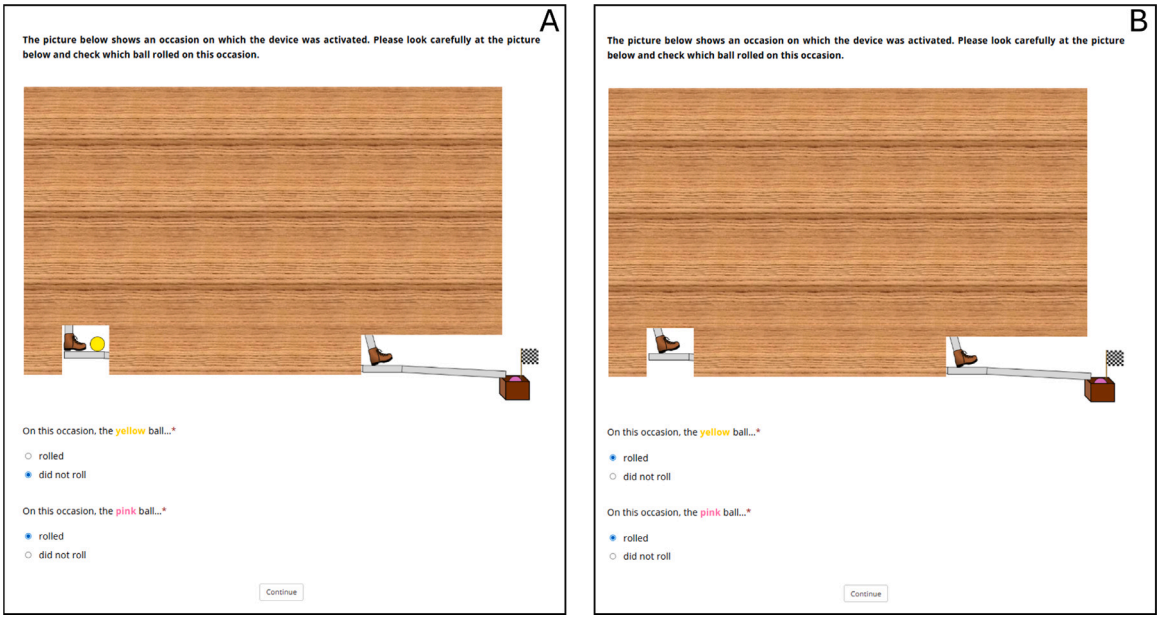

Fig. 8. Illustration of two learning trial screens in Experiment 1. A: a trial in which $C = 0$ and $E = 1$. B: a trial in which $C = 1$ and $E = 1$.

are you that the device shown above is one in which the rolling of the yellow ball can cause the rolling of the pink ball?" Responses were provided on a continuous slider recording values between 0 and 100 with a resolution of 1. The endpoints were labeled "Rolling of the yellow ball CANNOT cause rolling of the pink ball" (0) and "Rolling of the yellow ball CAN cause rolling of the pink ball (100);" the midpoint: "Uncertain" (50).

Participants were next informed about the learning phase: they would observe eight static pictures of the activated device, presented trial by trial. Participants in the transparent conditions were informed that the device's inside would not be visible during this phase. Two example learning screens are shown in Fig. 8. To support learning and establish a performance criterion, participants indicated on each trial which balls rolled, using radio buttons (cf. Fig. 8). Participants making errors on more than one of the eight trials were excluded.

After the learning phase, participants responded again to the previously used causal structure query to assess their structure posterior, followed by demographic questions, an opportunity to report technical problems, and a brief debriefing screen.

5.2. Results and discussion

The results are shown in Fig. 9. Individual causal structure ratings are shown in Fig. 9A; corresponding means in Fig. 9B. The experimental manipulation had the intended effect. Participants in the opaque conditions gave relatively high posterior causal ratings; they behaved as if they were experiencing a causal illusion. The pattern is consistent with

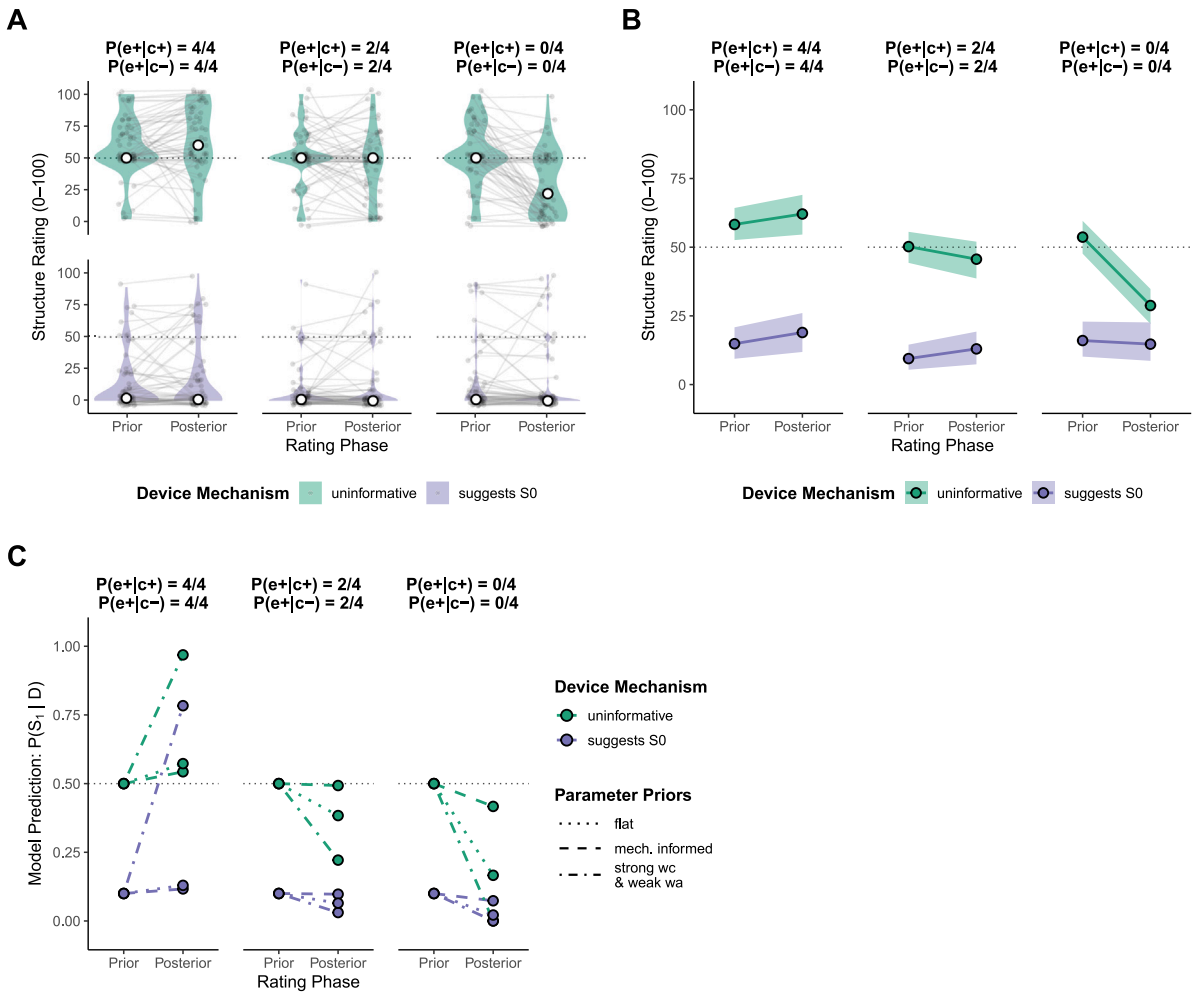

Fig. 9. Results of Experiment 1. A: Participants' prior and posterior causal structure ratings (jittered). Violin plots show their distribution; white dots show medians. B: Mean causal structure ratings. Ribbons show 95% CIs of the means. C: SI Model predictions based on different S_1 and parameter priors. The S_1 prior used in the uninformative mechanism condition was 0.5; in the transparent condition suggesting S_0, it was set to 0.10.

structure priors close to 0.5. Participants in the transparent mechanism conditions, by contrast, gave posterior ratings close to zero.[7] According to the SI model, the absence of a "causal illusion" in this condition reflects participants having entered the learning phase with a low S_1 prior — a prior that in light of consistent data does not need to be updated. Crucially, this pattern goes against the illusory causation hypothesis, according to which participants should have displayed an increased tendency to form a causal illusion the higher the outcome density is. Unlike participants in the transparent mechanism condition, participants in the opaque conditions were influenced by changes in $P(e^+ \mid c^-)$, that is, they were showing the outcome-density effect: the higher $P(e^+ \mid c^-)$, the higher their posterior ratings. Crucially, under the Bayesian learning view, this does not reflect the formation of a bias-based causal illusion; rather this effect is expected because higher levels of $P(e^+ \mid c^-)$ mean less evidence *against* the existence of a causal link.

These observations were statistically corroborated by a mixed ANOVA with causal rating type (CRT) as a within-subjects factor and contingency (C) and device mechanism (DM) as between-subjects factors, conducted with R's `afex` package (Singmann et al., 2022). Results are shown in Table 1. There was a significant main effect of DM, reflecting overall higher causal ratings in the opaque than in the transparent conditions. DM explained far more variance than any other factor (η_p^2 = .386), confirming that the prior causal belief substantially influences post-learning causal ratings. There was also a significant main effect of CRT, indicating lower posterior than prior ratings overall. Critically, this effect was moderated by DM, as shown by a significant CRT × DM interaction: the reduction from prior to posterior was driven by participants in the opaque conditions, especially by those in the (0/4)(0/4) condition in which the effect never occurred. This condition provided the strongest evidence against the existence of a causal link, thus prompting the strongest downward adjustments in the ratings. In the transparent conditions, participants' prior ratings were already low, which left little space for additional downward adjustment.

There was also a significant C × CRT interaction, reflecting differential updating across contingency conditions. This interaction further

Table 1
Results of Experiment 1: Type III analysis of variance for causal rating.

Effect	df	MSE	F	p	η_p^2
Device mechanism (DM)	1, 354	1005.96	222.65	< .001	.386
Contingency (C)	2, 354	1005.96	7.44	< .001	.040
DM × C	2, 354	1005.96	4.62	.011	.025
Causal rating type (CRT)	1, 354	290.31	6.45	.012	.018
DM × CRT	1, 354	290.31	17.62	< .001	.047
C × CRT	2, 354	290.31	16.15	< .001	.084
DM × C × CRT	2, 354	290.31	7.28	< .001	.040

[7] This near-zero effect was obtained despite the use of a unidirectional rating slider, which has previously been found to inflate the magnitude of the effect (see Ng et al., 2024).

interacted with DM, as indicated by a significant DM × C × CRT interaction: the degree to which causal ratings decreased from prior to posterior across contingency conditions (i.e., changes in outcome density) was stronger in the opaque than in the transparent conditions. This pattern matches the predictions of the SI model.

5.2.1. Model fits

The mean causal ratings were in line with the qualitative predictions of the SI model. As an exploratory step, we compared participants' mean ratings with the predictions of different quantitative versions of the SI model. These predictions are shown in Fig. 9C.

For the opaque mechanism condition, we set the S_1 structure prior to 0.5. For the transparent condition suggesting the absence of a causal link, we set it to a low value of 0.10. We then generated predictions under three sets of parameter prior distributions.

The first version used the default flat Beta$(1, 1)$ distributions over w_c and w_a (cf. Fig. 3), labeled "flat" in Fig. 9C (dotted lines).

The second version used a set of strength priors representing a plausible alternative default for learning about artifacts: a w_c prior favoring a strong target cause and a w_a prior favoring a weak alternative (cf. Fig. 4), labeled "strong w_c & weak w_a" (dot-dashed lines).

The third version used parameter priors matched to the specific mechanism components shown in the device illustrations. The wall between yellow and pink ball was assumed to elicit a w_c prior favoring very low values, implemented with Beta$(1, 10)$ (cf. Fig. 5C). The w_a prior varied by contingency condition, reflecting the different alternative-cause components shown in the stimuli: Beta$(80, 80)$ for the $P(e^+|c^-) = 0.5$ condition, Beta$(1, 10)$ for $P(e^+|c^-) = 0$, and Beta$(40, 5)$ for $P(e^+|c^-) = 1.0$. This version is labeled "mech. informed" (dashed lines).

Comparing model versions by RMSE and the correlation between predicted and observed means, the default flat-prior model best aligned with participants' mean ratings. In the opaque condition, it yielded the lowest RMSE (7.13) and the highest correlation coefficient ($r = .973$). In the transparent condition, the flat-prior model ($RMSE = 7.02$, $r = .282$) and the mechanism-specific prior model ($RMSE = 5.44$, $r = .329$) performed comparably. It is thus possible that at least some participants in the transparent condition were influenced by both structural and parametric information conveyed by the mechanism illustrations. The model with a strong w_c and a weak w_a prior fitted the means worst, even in the opaque condition in which such a prior set might have been a plausible alternative default ($RMSE = 21.10$, $r = .905$).

5.2.2. Interindividual differences

We additionally examined participants' causal ratings at the individual level. As Fig. 9A shows, not all participants displayed the pattern described by the means. We therefore classified participants according to changes between their prior ($P(S_1)$) and posterior judgments ($P(S_1|D)$); results are shown in Fig. 10. In the opaque conditions, the proportion of participants whose posterior ratings were lower than their prior ratings (purple bars) became smaller as $P(e^+ \mid c^-)$ (i.e., outcome-density) increased, consistent with the aggregate analyses. According to the illusory causation view, this is because higher levels of outcome density make learners more prone to perceive an "illusory" causal link. Under the Bayesian learning hypothesis, the reason for this pattern is that the higher the outcome-density under zero-contingency data is, the less diagnostic these data are for the absence of a causal relation (i.e., for S_0). Here fewer participants are according to the Bayesian account expected to downward-adjust their prior causal belief in the existence of a causal link.

In the transparent conditions, the largest cluster across all contingency conditions was the one in which participants gave identical prior and posterior ratings. Unlike in the opaque conditions, most participants kept their low prior belief in S_1 even if they encountered a high outcome density. This behavior is inconsistent with the illusory causation view, according to which participants' inclination to perceive an "illusory" causal link should have increased with increasing levels of outcome density. The pattern is, however, consistent with the Bayesian view: encountering learning data that is consistent with a low prior belief in the existence of a causal link should not elicit much updating.

Yet, it can also be seen that some participants gave higher posterior than prior ratings, even in the (2/4)(2/4) condition, where the data provide evidence against S_1 according to the SI model (note that a small increase is even expected under ceiling conditions). This subgroup of participants is clearly visible in Fig. 9A, where it can be seen that some of these participants increased their ratings substantially. This suggests that a minority of participants may have exhibited non-normative "causal illusions" even after viewing mechanism information suggesting the absence of a link. This may have happened because they engaged in differential weighting of trial types (i.e., overweighted $[c^+, e^+]$ trials), one of the explanations for the formation of genuine causal illusions (see Perales et al., 2017).

6. Experiment 2

A potential limitation of Experiment 1 is that participants in the transparent conditions viewed the device mechanism twice: once during the prior rating and again on the posterior test screen. This provided a second opportunity to analyze the mechanism after the learning phase. Although participants were required to report observed events on each trial during learning, and the analyses were restricted to those who demonstrated high accuracy, this repeated exposure to the mechanism illustration may have attenuated or overridden the influence of the learning data. One could argue that this is what kept participants in the transparent conditions from showing a causal illusion in their posterior ratings.

Moreover, even without a second presentation of the mechanism, the visual prior manipulation may have been so salient that it caused participants to discount any subsequent empirical evidence in favor of their initial structural intuitions. This suggests an alternative explanation: the near-zero causal ratings after learning in the transparent conditions may not have resulted from data confirming low prior causal beliefs but rather from insufficient attention to the data.

Experiment 2 was designed to rule out both hypotheses. First, participants in the transparent conditions were shown the target device's mechanism only once, on a separate screen prior to the learning phase. They were informed that the device would remain covered for the rest of the experiment, and that this screen provided the sole opportunity to study its mechanism. The test screens in the transparent conditions subsequently showed a covered device, visually identical to those in the opaque conditions (cf. Fig. 7A).

To test whether participants in Experiment 1 had simply ignored mechanism-congruent learning data, we introduced conditions designed to elicit causal priors *incongruent* with the data. If participants engage in Bayesian updating rather than merely discounting evidence, their posterior judgments should reflect the mismatch between priors and data.

Specifically, we added a positive-contingency condition (see the left panel in Fig. 3B) in which the observations provided evidence for a causal link. This should elicit upward adjustments in posterior ratings, particularly for participants who were initially undecided (opaque conditions) or held low S_1 priors (transparent low-prior conditions). We also included an additional high-prior transparent condition in which the physical barrier between the yellow and pink balls was removed. Unlike low-prior participants, participants in this condition should perceive zero-contingency data as evidence against their initial causal belief. Bayesian integration predicts a downward revision of posterior ratings (cf. Fig. 3B).

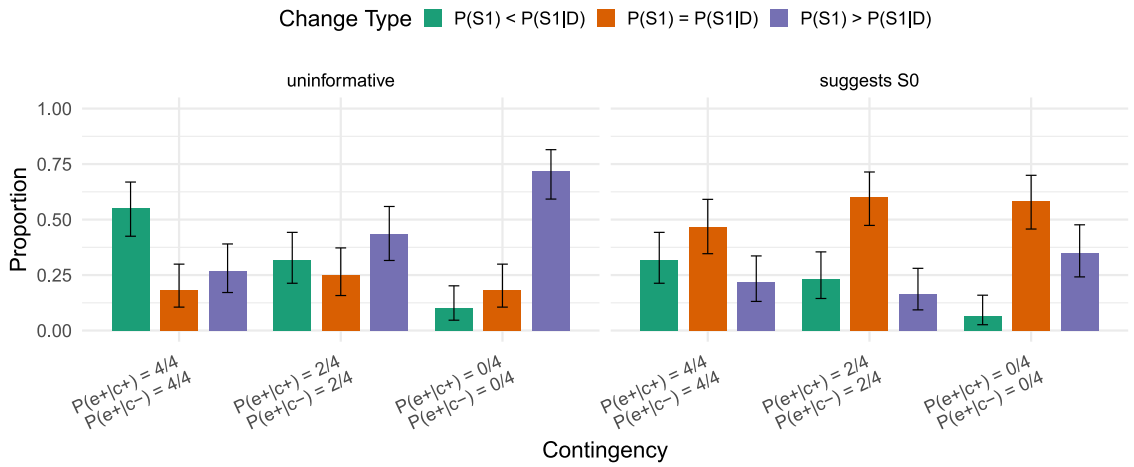

Fig. 10. Results of Experiment 1: Proportions of observed changes from prior to posterior structure judgment. Error bars show 95% Wilson CIs.

6.1. Methods

6.1.1. Participants

We recruited $N = 665$ participants via www.prolific.com. Inclusion criteria were identical to Experiment 1. Two participants were excluded because they did not confirm participation via a PC or laptop. An additional 123 participants (18.55%) were excluded for not meeting the pre-registered learning-phase performance criterion.[8] The final analysis sample consisted of $N = 540$ participants (251 male, 282 female, 5 non-binary, 5 preferred not to say; $M_{age} = 38$ years, range 18–78 years). As pre-registered, data collection was terminated once all conditions included $n = 60$ participants after exclusions, matching the per-condition sample size of Experiment 1.

6.1.2. Design, materials, and procedure

The experiment employed a 3 (device mechanism: opaque vs. transparent with wall [suggesting S_0] vs. transparent without wall [suggesting S_1]; between-subjects) × 3 (contingency: (4/0)(0/4) vs. (2/4)(2/4) vs. (0/4)(0/4); between-subjects) × 2 (causal rating type: prior vs. posterior; within-subjects) mixed design.

Materials and procedure were similar to the ones in Experiment 1. The key procedural difference was that mechanism illustrations in the transparent conditions were shown only once on a dedicated screen after the general introduction and prior to the contingency learning phase. Participants were told that this screen provided the sole opportunity to study the target device's interior. Unlike in Experiment 1, the test screens in the transparent conditions displayed a covered device and were thus visually identical to those in the opaque conditions. Another difference from Experiment 1 was the additional transparent condition in which the wall between the yellow and the pink ball was removed, which was intended to elicit high S_1 prior ratings.

6.2. Results and discussion

The results are shown in Fig. 11. Individual ratings are in Fig. 11A; means in Fig. 11B. In the two zero-contingency conditions, the results replicate those of Experiment 1, although mean ratings were overall somewhat higher: participants in the transparent low-prior conditions (purple) gave lower prior ratings than those in the opaque conditions (green), and correspondingly lower posterior ratings, again attenuating the outcome-density effect.

The novel high-prior transparent condition yielded two clear findings. First, participants in these conditions gave higher prior ratings than those in all other conditions. Second, the learning data had a clear impact on posterior ratings, particularly when data conflicted with the high prior: participants adjusted their posterior ratings downward. Nevertheless, because of their high initial belief in S_1, their posterior ratings remained higher than those of participants in the other conditions — replicating the pattern that previous articles have labeled "causal illusions." These findings speak against the hypothesis that our visual mechanism manipulation simply prevented participants from processing the data.

The results from the positive-contingency condition further support this conclusion: low-prior participants tended to upward-adjust their causal ratings in light of evidence for a causal link, confirming that they did incorporate the learning data.

The statistical analyses corroborated these interpretations. Using R's `lme4` package (Bates et al., 2015), we fitted a mixed-effects model with device mechanism, contingency, and causal rating type as fixed effects, and participant ID as random effect. Results are shown in Table 2 as a Type 3 ANOVA table.

Table 2
Results of Experiment 2: Type III analysis of variance for causal rating.

Effect	df	MSE	F	p	η_p^2
Device mechanism (DM)	2, 531	1391.02	104.47	< .001	.282
Contingency (C)	2, 531	1391.02	25.15	< .001	.087
DM × C	4, 531	1391.02	0.85	.493	.006
Causal rating type (CRT)	1, 531	472.53	38.70	< .001	.068
DM × CRT	2, 531	472.53	20.25	< .001	.071
C × CRT	2, 531	472.53	46.40	< .001	.149
DM × C × CRT	4, 531	472.53	2.43	.040	.018

There was a significant main effect of DM. Contrast tests showed that ratings in the high-prior conditions were overall higher than in the uninformative (opaque) conditions, $t(531) = 5.858$, $p < .001$, and ratings in the uninformative conditions were overall higher than in the low-prior conditions, $t(531) = -8.515$, $p < .001$. There was also a significant main effect of CRT: posterior ratings were lower than prior ratings overall, $t(531) = 6.221$, $p < .001$, but this effect depended on contingency condition, as indicated by a significant CRT × C interaction. The interaction arose because the zero-contingency conditions tended to elicit downward adjustments while the positive-contingency condition tended to elicit upward adjustments. These upward adjustments were also observed in low-prior participants, $t(531) = -2.368$, $p = .018$, confirming that they paid attention to and incorporated the learning data. The significant three-way DM × C × CRT interaction further shows that the degree to which causal ratings changed from prior to posterior across contingency conditions differed between mechanism conditions, in a pattern consistent with Bayesian updating.

[8] The pre-registered criterion was slightly less strict than in Experiment 1: participants with at least 6/8 correct trials were included.

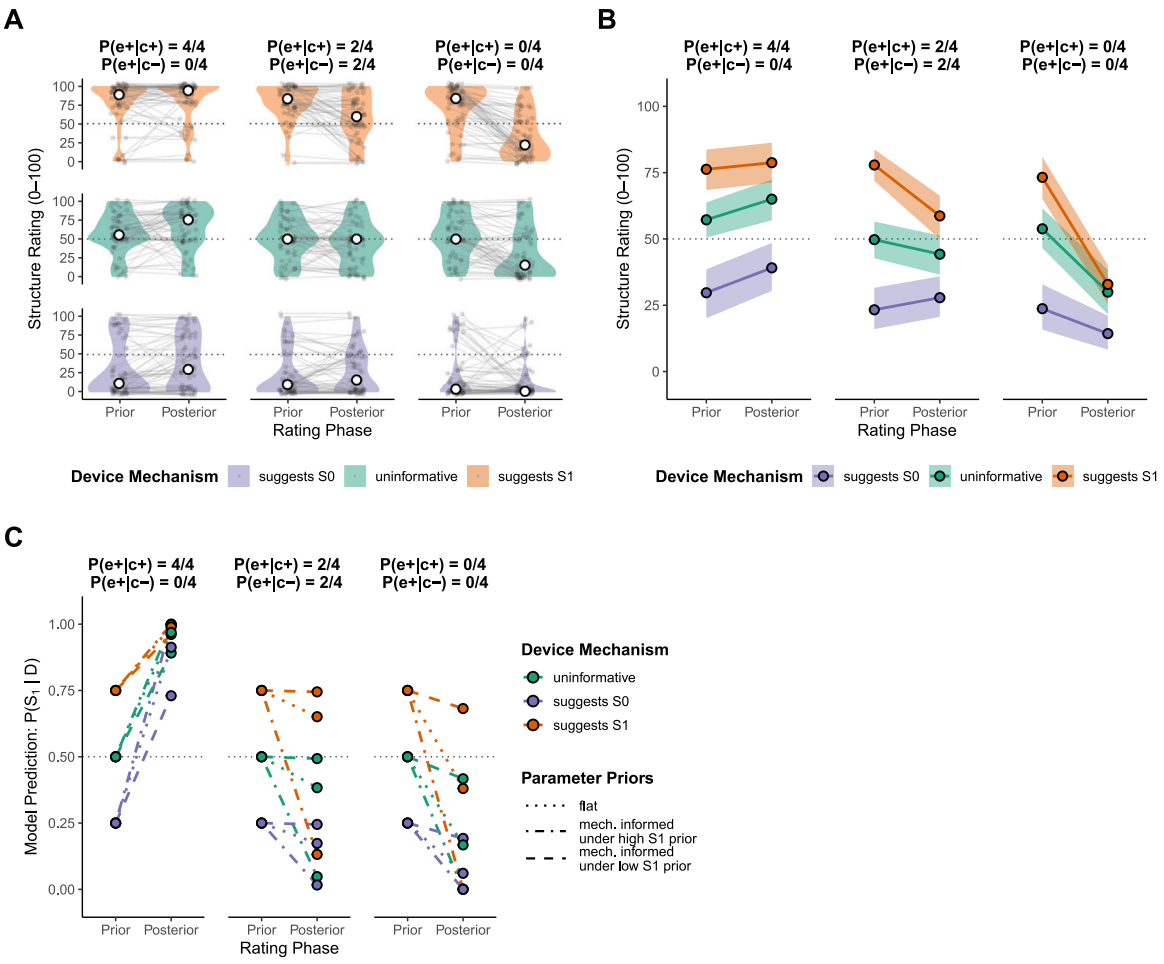

Fig. 11. Results of Experiment 2. A: Participants' prior and posterior causal structure ratings (jittered). Violin plots show their distribution; white dots show medians. B: Mean causal structure ratings. Ribbons show 95% CIs of the means. C: SI Model predictions. The S_1 prior was 0.5 in the uninformative condition, 0.25 in the transparent condition suggesting S_0, and 0.75 in the transparent condition suggesting S_1.

6.2.1. *Model fits*

Given that Experiment 2 included a novel positive-contingency condition and a novel high-prior transparent condition, we again conducted exploratory model fits. Predictions are shown in Fig. 11C. Based on the observed mean prior ratings, we set the S_1 priors to 0.25, 0.5, or 0.75 for the three mechanism conditions. These values were informed directly by participants' mean prior ratings. As for the parameter priors, we implemented one version of the model, like in Experiment 1, in which we used a flat Beta$(1, 1)$ over the parameters, b_c, w_c, and w_a (keeping w_c at 0 under S_0). We also implemented two model versions with parameter priors that are more or less plausible given the observed mechanism in the transparent mechanism conditions. In one version (labeled "mech. informed under high S_1 prior" in Fig. 11C), we used a Beta$(40, 5)$ distribution over w_c (cf. Fig. 4C), which reflects a prior belief in a high causal strength of C. If participants are sensitive to the parameter information conveyed by the visual mechanism illustrations, then this model may be most accurate in the transparent condition, where no wall is positioned on the path from the yellow ball to the pink ball (i.e., the high S_1 prior condition). In a second version (labeled "mech. informed under low S_1 prior" in Fig. 11C), we used a Beta$(40, 5)$ distribution over w_c (cf. Fig. 5C), reflecting a prior belief in low causal strength of C. If participants are sensitive to the parameter information conveyed by the visual mechanism illustrations, then this model may be most accurate in the transparent condition in which a wall is positioned on the path from the yellow to the pink ball (i.e., the low S_1 prior condition). For both model versions, the prior distribution over w_a was either a Beta$(1, 10)$ or a Beta$(80, 80)$, depending on whether the component of the mechanism representing the influence of the background cause A suggested a value of w_a close to 0 or close to 0.50.

As in Experiment 1, model fit was assessed by RMSE and correlation coefficient. In the opaque condition, the flat-prior model performed best ($RMSE = 14.70$, $r = .909$). In the transparent low-S_1-prior condition, the mechanism-specific model (labeled "mech. informed under low S_1 prior" in Fig. 11C) and the flat-prior model performed similarly ($RMSE = 14.20$, $r = .820$ and $RMSE = 22.10$, $r = .858$, respectively). In the transparent high-prior condition, the flat-prior model had a smaller RMSE than the mechanism-specific model (labeled "mech. informed under high S_1 prior" in Fig. 11C; 9.02 vs. 24.50), but the correlations between model predictions and participants' ratings were similar for both model versions ($r = .907$ vs. $r = .916$). In sum, the overall good fit of the flat prior model suggests that participants primarily focused on structural information, though some participants may also have formed mechanism-based assumptions about causal parameters. This suggests that different subgroups of participants may have relied on different reasoning strategies, which is why we decided to look at interindividual differences next, as in Experiment 1.

6.2.2. *Interindividual differences*

Individual ratings, which are shown in Fig. 11A, again revealed substantial interindividual variance. A subgroup of participants did not update their ratings at all. However, an inspection of Fig. 11A shows that this pattern was most common among participants who expressed extreme prior beliefs. In this case, little to no updating is actually predicted by the SI model (cf. Fig. 3).

Another interesting pattern visible in Fig. 11A is the substantial interindividual variability in the magnitude of prior-to-posterior belief change. In the (0/4)(0/4) condition in particular, a substantial number of participants in the transparent mechanism condition conveying a high S_1 prior showed larger downward adjustments than those in the opaque (uninformative) mechanism condition. This produced a pattern of mean rating differences (cf. Fig. 11B) in which the prior-posterior difference was greater in the transparent high-S_1 prior condition than in the opaque condition. At first glance, this appears to contradict the Bayesian updating principle discussed in the theoretical section, according to which a higher S_1 prior creates resistance to movement when disconfirming evidence is encountered — which would lead one to expect a *smaller* prior-posterior shift in the transparent high-S_1 prior condition than in the opaque condition. However, this is true only under the assumption that participants in both conditions relied on the *same* prior parameter distributions. If at least a subset of participants in the transparent high-S_1 prior condition relied on mechanism-informed rather than flat parameter priors, stronger downward updating in that condition is in fact predicted by the model. Comparing the SI model predictions for the (0/4)(0/4) condition shown in Fig. 11C with the individual ratings in Fig. 11A, the behavior of participants who showed the largest downward adjustments in the transparent high-S_1 prior condition closely approximates what the SI model version ("mech. informed under high S_1 prior") using mechanism-informed parameter priors that align with the mechanism information conveyed by the stimuli predicts.

There were, however, also participants whose behavior clearly seems to contradict rational Bayesian updating. For example, Fig. 11A shows that a number of participants in the zero-contingency conditions, especially the (2/4)(2/4) condition, upward-adjusted their posterior ratings. This is not predicted by the model, irrespective of the specific priors used to compute the predictions. This pattern of behavior may therefore indicate the occurrence of genuine causal illusions in at least some individuals. Although the mean difference was not significant in this condition ($t(531) = 1.151$, $p = .250$), this replicates the pattern already observed in Experiment 1.

Taken together, the results of Experiment 2 are overall well accounted for by a Bayesian view of causal induction:learners' posterior causal beliefs — including the magnitude of apparent "causal illusions" — depend on (1) participants' prior causal beliefs and (2) the degree to which the encountered evidence speaks for or against these beliefs.

An interesting open question emerging from Experiment 2 is how participants understand situations in which there is a mismatch between prior information and subsequent contingency information. How do participants make sense, for example, of a mechanism that seems to rule out a positive contingency in light of later presented positive contingencies? One possibility is that some participants do not pay sufficient attention to the setup so that they used all the information they received despite the mismatch. Also, since we only showed static mechanisms, which required mental simulations to generate processes, there may be some uncertainty about possible paths of the balls. It is possible that some participants considered improbable situations, such as the cause ball jumping over the wall.

7. Experiment 3

Previous studies on apparent causal illusions under zero contingencies asked causal strength rather than structure queries, whereas we used structure queries in Experiments 1 and 2. We did so because we believe that the effect is fundamentally structural in nature (see also Griffiths & Tenenbaum, 2005; Ng et al., 2026). However, one might argue that our results might have differed if we had asked strength queries. Specifically, it may turn out that the manipulation of the prior will lose its influence. To rule out this possibility, we conducted a replication of our task using a strength query, following the formulation of Buehner et al. (2003).

We expected the manipulation of the prior to influence strength ratings for the same reason it should influence structure ratings: a structural belief can be expressed on a strength scale. A participant who believes in the absence of a causal link can indicate a strength of zero (or close to zero to express residual uncertainty), and one who is not convinced of the link's absence may report a value above zero.

7.1. Methods

7.1.1. Participants

We recruited $N = 580$ participants via www.prolific.com. Inclusion criteria were the same as in the previous experiments. One participant was excluded for neither confirming participation via a PC or laptop nor agreeing to pay attention. An additional 99 participants (17.10%) were excluded for failing to meet the pre-registered performance criterion used in Experiment 2. The final analysis sample consisted of $N = 480$ participants (210 male, 259 female, 10 non-binary, 1 preferred not to say; $M_{age} = 37$ years, range 18–76 years). Data collection was terminated once all conditions included $n = 60$ participants after exclusions.

7.1.2. Design, materials, and procedure

The experiment had a 2 (device mechanism: opaque vs. transparent with wall [suggesting S_0]; between-subjects) × 4 (contingency: (4/0)(0/4) vs. (4/4)(4/4) vs. (2/4)(2/4) vs. (0/4)(0/4); between-subjects) × 2 (causal rating type: prior vs. posterior; within-subjects) mixed design. The high-prior condition from Experiment 2 was dropped, but the positive-contingency condition was retained.

The materials and procedure were identical to the ones in Experiment 2, with one exception: participants rated the causal strengths rather than making judgments about causal structure. The strength query followed Buehner et al. (2003) and read: "In the present device above, how strongly does the rolling of the yellow ball cause the rolling of the pink ball?" Slider endpoints were labeled "Rolling of the yellow ball does not cause rolling of the pink ball at all"[9] and "Rolling of the yellow ball causes rolling of the pink ball all the time."

7.2. Results and discussion

The results, shown in Fig. 12, replicate our previous findings. In the transparent conditions, where participants saw a mechanism suggesting the absence of a causal link (S_0), both prior and posterior causal strength ratings were lower than in the opaque conditions. This confirms that the degree to which reasoners perceive a causal connection after learning depends on their prior belief in such a connection. This was now seen also in strength queries. The influence of the learning data also replicated: in the zero-contingency conditions, lower outcome density was associated with lower posterior ratings, consistent with data congruent with a low prior. The positive-contingency condition showed that participants correctly increased their ratings from prior to posterior probabilities when the data supported a causal relation; this held in the low-prior condition as well, confirming that participants who viewed the mechanism suggesting S_0 also paid attention to the learning data and incorporated them in their judgments.

A mixed-effects model with device mechanism, contingency, and causal rating type as fixed effects, and participant ID as random effect, yielded a main effect of device mechanism, $F(1, 472) = 102.34$, $p < .001$, confirming that ratings in the opaque conditions were overall higher than in the transparent conditions. There was also a main effect of contingency, $F(3, 472) = 6.92$, $p < .001$, and a significant contingency × causal rating type interaction, $F(3, 472) =$

[9] Although this endpoint reads like a structural judgment, it is appropriate for a strength scale: a causal strength of exactly zero is conceptually equivalent to the absence of a causal link.

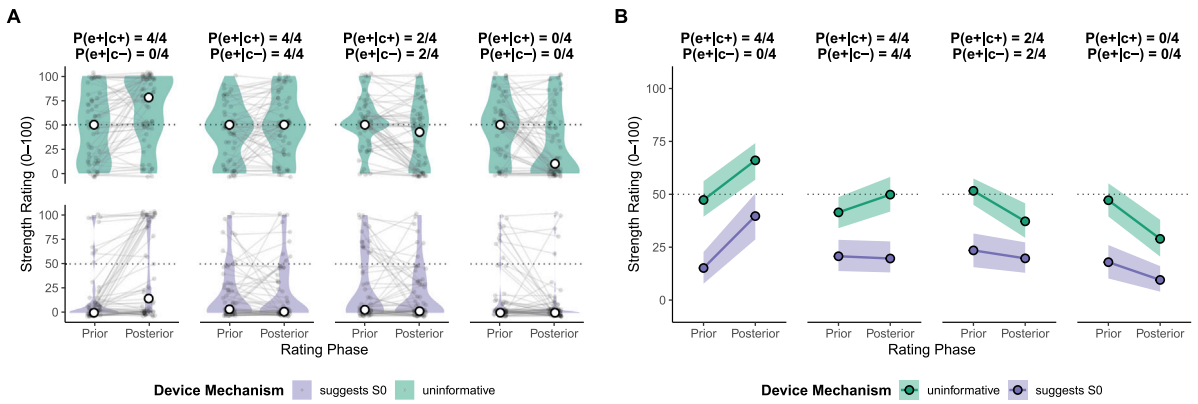

Fig. 12. Results of Experiment 3. A: Participants' prior and posterior causal strength ratings (jittered). Violin plots show their distribution; white dots show median ratings. B: Mean causal strength ratings. Ribbons show 95% CIs of the means.

28.55, $p < .001$, reflecting differential updating depending on whether data supported the presence or absence of a causal link. The three-way interaction between device mechanism, contingency, and causal rating type did not reach significance, $F(3, 472) = 2.54$, $p = .056$, thus only suggesting a trend in the expected direction. As Fig. 12 shows, participants in both prior conditions showed similar prior-to-posterior shifts across contingency conditions; the only qualitatively different pattern appeared in the ceiling contingency, where participants in the uninformative prior condition slightly increased their prior ratings. This replicates a similar finding from Experiment 1 consistent with SI model predictions.

8. Experiment 4

In the previous experiments, we manipulated causal structure priors using visual depictions of device mechanisms. Experiment 4 retains artificial devices as learning material, but contrasts priors conveyed by visual mechanism information with priors conveyed by statistical frequency information, an alternative way of prior manipulation involving information about the statistical prevalence of causal versus non-causal mechanisms in the devices (as discussed in Section 3.2). Theoretically, the source of a structural prior should be irrelevant to how it is updated in light of new evidence: a prior of $P(S_1) = 0.10$ should influence the posterior identically whether it stems from a visually conveyed intuition or base-rate information.

However, our visual mechanism illustrations also contain information about the size of causal parameters, and may therefore shape expectations not only about the existence of a causal link (S_1 vs. S_0) but also about the strength of the links of the given structure. For example, the tube systems depicted in the mechanism illustrations (cf. Fig. 6) suggest specific event probabilities, such as the probability of the pink ball rolling in the absence of a rolling of the yellow ball, $P(e^+|c^-)$. The right tube systems of the devices depicted in Fig. 6B and C, for example, suggest that $P(e^+|c^-)$ is higher in the device shown in Fig. 6B than in the one shown in Fig. 6C. The model fits in Experiments 1 and 2 indicated that the flat-prior model performed well, suggesting that participants may have focused primarily on the structural information in the stimuli and not on the size of the causal parameters. This is in line with the view that structure matters more than strength. However, we also found in Experiment 2 that a number of participants seemed to have incorporated the visual information about the causal parameters, which then influenced their judgments accordingly. For this reason, we wanted to conduct a more direct test of whether a visual mechanism manipulation that targets both structure and parameter priors produces substantially different updating behavior on average than a statistical manipulation targeting structure priors alone.

8.1. Methods

8.1.1. Participants

We recruited $N = 541$ participants via www.prolific.com. Inclusion criteria were the same as in the previous experiments. Two participants were excluded for not confirming participation via a PC or laptop. An additional 59 participants (10.95%) were excluded for not meeting the pre-registered performance criterion we also used in Experiments 2 and 3. The final analysis sample consisted of $N = 480$ participants (206 male, 268 female, 5 non-binary, 1 preferred not to say; $M_{age} = 37$ years, range 18–76 years). Data collection was terminated once all conditions included $n = 60$ participants after exclusions.

8.1.2. Design, materials, and procedure

We tested two zero-contingency data sets, (4/4)(4/4) and (2/4)(2/4), and aimed to induce two levels of prior belief: a weak belief in S_1 versus a strong belief in S_1. These priors were conveyed either by visual mechanism information or by statistical information about the frequency of causal and non-causal devices. The study thus employed a 2 (target prior level: low [suggesting S_0] vs. high [suggesting S_1]; between-subjects) × 2 (prior information type: mechanistic vs. statistical; between-subjects) × 2 (contingency: (4/4)(4/4) vs. (2/4)(2/4); between-subjects) × 2 (causal rating type: prior vs. posterior; within-subjects) mixed design.

Materials and procedure were largely identical to the ones in previous experiments. In the statistical prior condition, participants saw after the general instructions a screen on which they learned that scientists had constructed 10 devices in total. This was accompanied by a picture showing the ten devices, arranged in a grid with their insides covered. The description emphasized that all devices look identical from the outside but may differ internally. Participants were then informed about the number of causal devices (in which yellow ball movement can cause pink ball movement) and non-causal devices (in which pink ball movement always results from an alternative internal mechanism). Depending on the condition manipulating the prior level, participants either learned that one of the ten devices was causal (low prior, 10%) or that nine out of ten were causal (high prior, 90%).

The visual mechanism-information conditions mirrored those of the previous experiments. Rather than statistical base rates, participants viewed a depiction of the target device's internal mechanism, showing either a wall obstructing the path between the yellow and pink balls (low-prior condition) or an open path (high-prior condition). These illustrations were slightly modified for Experiment 4 (see Fig. 13), shifting the alternative trigger further to the left, away from the pink ball. This modification was necessitated by the introduction of a high-prior condition paired with a ceiling contingency — a combination not previously tested (bottom left illustration in Fig. 13). Because the original proximity of the alternative trigger to the pink ball suggested

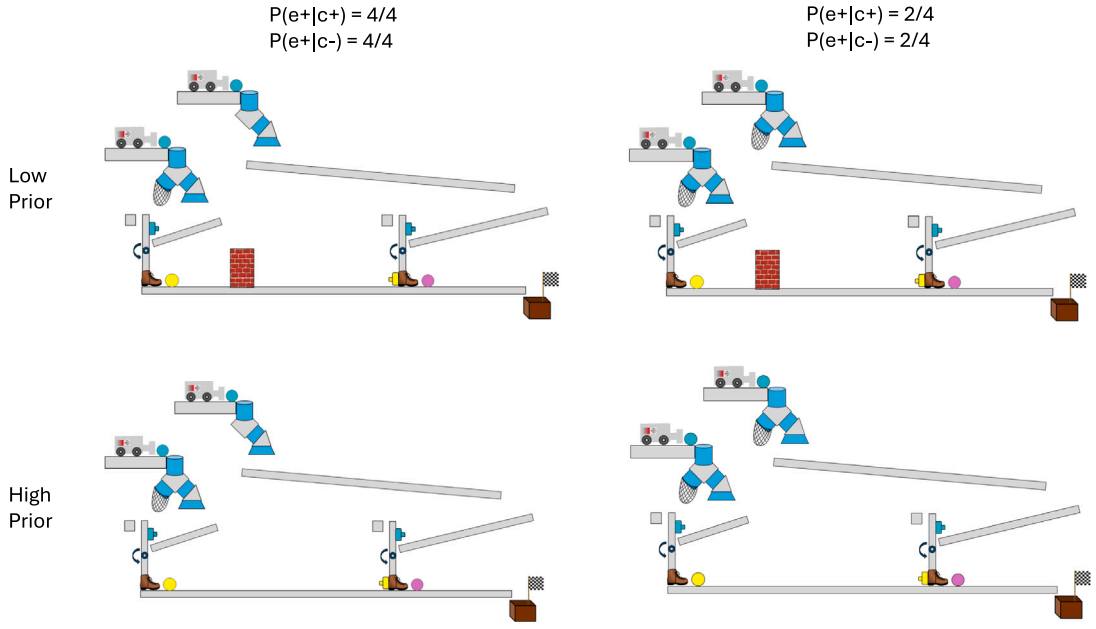

Fig. 13. Illustration of the mechanism-information stimuli used in Experiment 4.

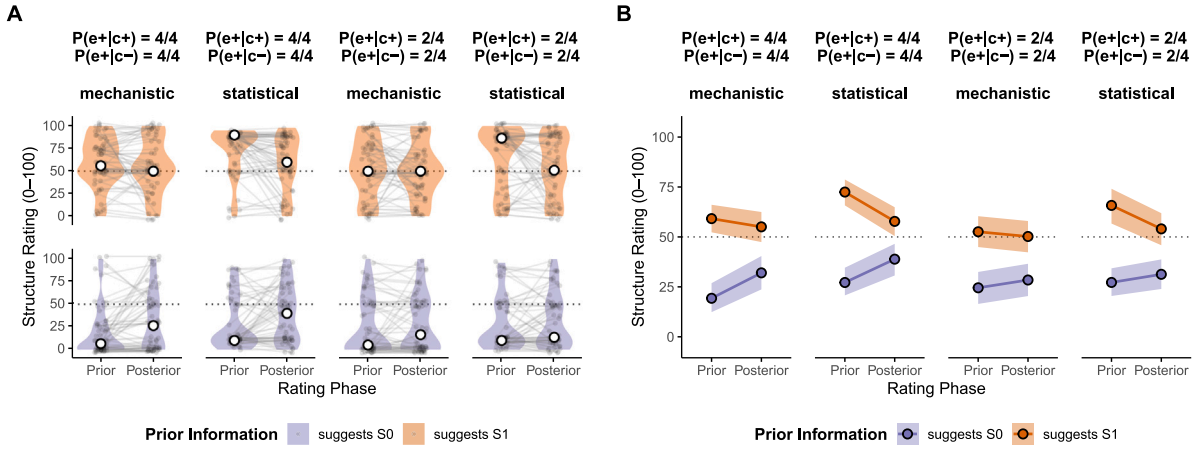

Fig. 14. Results of Experiment 4. A: Participants' prior and posterior causal structure ratings (jittered). Violin plots show their distribution; white dots show median ratings. B: Mean causal structure ratings. Ribbons show 95% CIs of the means.

it might preempt the yellow ball's influence on occasions on which both move towards the pink ball (Stephan et al., 2020; Stephan & Waldmann, 2018), its presence in a *causal* device that generates ceiling data (in which C and A always act to generate E) risked leading participants to infer that the yellow ball's causing the pink ball to move is generally pre-empted by the alternative trigger. Moving the trigger further left increases its expected travel time, making the yellow ball appear faster on occasions when both move toward the pink ball.

For the test queries, we returned to causal structure queries, with a modified formulation better suited to both types of information about the priors. The prior and posterior query read: "What's the probability that the present device above is one in which the rolling of the yellow ball can cause the rolling of the pink ball (i.e., a causal device)?" Slider endpoints were "Certainly not causal" and "Certainly causal."

8.2. Results and discussion

The results are shown in Fig. 14. The statistical prior manipulation was effective: participants in these conditions gave higher prior causal ratings when the statistical information suggested S_1 than when it suggested S_0. In the visual mechanism-information conditions, prior ratings similarly differed depending on whether the mechanism illustration suggested the presence or absence of a causal link. However, the high-prior condition in the visual mechanism-information group appeared less effective than in Experiment 2, suggesting that the modified position of the alternative cause, contrary to what we intended, introduced more uncertainty. Notably, this also occurred in the (2/4)(2/4) condition, where causal preemption was less of a concern. As a result, mean prior ratings in the high-prior mechanism-information conditions were lower than in the corresponding statistical conditions.

Overall, however, the results corroborate again the Bayesian interpretation of the outcome-density effect: the degree to which learners perceive a causal connection after learning is influenced by their prior causal belief, with higher priors yielding higher posteriors.

A mixed-effects model with the size of the prior, prior information type, contingency, and causal rating type as fixed effects, and participant ID as random effect, yielded a main effect of the prior level, $F(1, 472) = 147.80$, $p < .001$, confirming the effectiveness of the manipulation of the prior and accounting for the most variance ($\eta_p^2 = .238$). There was also a main effect of prior information type,

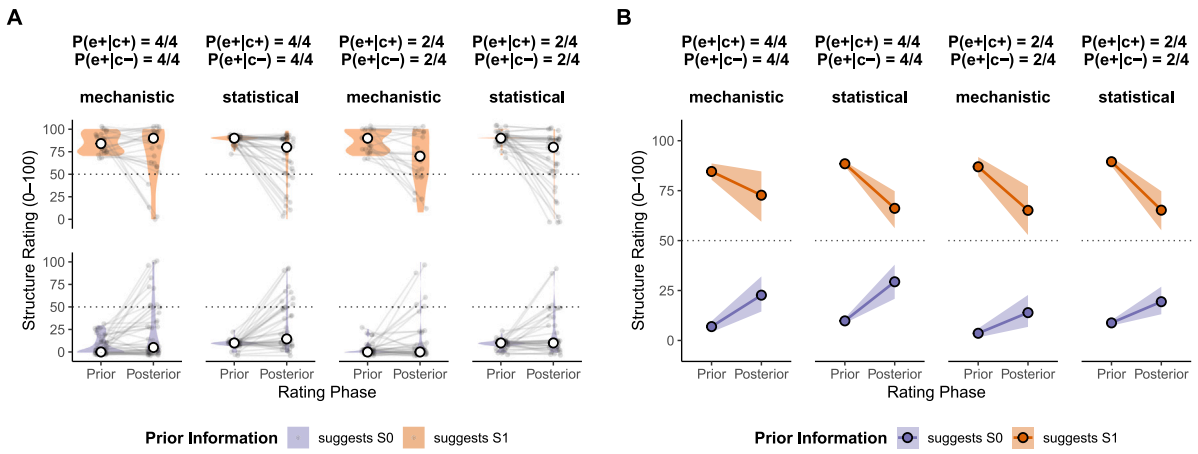

Fig. 15. Results of Experiment 4 (subgroup data, $N = 293$). A: Participants' prior and posterior causal structure ratings (jittered). Violin plots show their distribution; white dots show median ratings. B: Mean causal structure ratings. Ribbons show 95% CIs of the means.

$F(1, 472) = 7.41$, $p = .006$, $\eta_p^2 = .015$, reflecting overall higher ratings in the statistical than in the visual mechanism-information conditions. A significant interaction between the prior level and causal rating type, $F(1, 472) = 38.70$, $p < .001$, $\eta_p^2 = .076$, showed that updating from prior to posterior was modulated by the level of prior belief: high-prior participants tended to adjust downward, while low-prior participants tended to adjust upward. A significant three-way interaction between the level of the prior, causal rating type, and contingency, $F(1, 472) = 4.05$, $p = .009$, further confirmed that the degree of updating was influenced by the learning data.

To test directly whether the two types of prior information led to different updating behaviors, we conducted an exploratory analysis on the subset of participants who had given clearly high or low prior ratings ($N = 293$). We analyzed all participants in the high-prior conditions with prior ratings of at least 70, and all participants in the low-prior conditions with ratings no higher than 30. The results of this homogeneous subsample with respect to the prior are shown in Fig. 15. Participants who received visual mechanism information and those who received statistical information behaved very similarly, corroborating the model fit results from our earlier experiments. What primarily drove participants' behavior was the structural information conveyed by the stimuli, regardless of whether it was presented visually or statistically. The results also show that the Bayesian explanation of the outcome-density effect holds across different types of prior manipulation.

A mixed-effects model fitted to this subsample yielded only a significant main effect of level of the prior, $F(1, 285) = 1162.05$, $p < .001$, and a significant interaction between the level of the prior and the causal rating type, $F(1, 285) = 98.99$, $p < .001$. The factor prior information type yielded neither a significant main effect nor a significant interactions with any other factor.

In a supplementary study following up on the question of whether participants look more at the structural rather than the parametric information conveyed by the mechanism stimuli, we compared two visual conditions: one in which the whole mechanism was visible (structure and parameter information) and one in which only the path between the yellow and the pink ball was uncovered (structure information). The results in both conditions were nearly identical, providing direct evidence that participants mostly focused on structural information. A summary of the supplementary study is provided in the repository.

9. General discussion

The finding of non-zero causal ratings after zero-contingency learning data has long been treated as a paradigm case of irrational causal reasoning — a "causal illusion" arising when learners confuse coincidence with causation. This bias perspective is so prevalent and taken for granted that some even claimed that given zero-contingency learning data "[...] independently of the norm taken as reference, any non-zero causal judgments should be considered as biased" (Perales et al., 2017, pp. 36–37). Based on (normative) accounts of Bayesian causal induction and based on four experiments reported here, we challenge this characterization. By successfully manipulating causal structure priors and showing that these priors influenced post-learning causal judgments in the direction predicted by Bayesian updating, we provide direct empirical support for the rational Bayesian account of causal-illusions under zero-contingencies, and the influence exerted on them by variations in outcome-density.

We believe that one reason why this rational explanation, although first proposed roughly twenty years ago (Griffiths & Tenenbaum, 2005), has received little empirical attention is that it has rarely been the focus of direct investigation. An exception is a recent study by Ng et al. (2026), who assessed participants' causal priors before presenting learning data. They found that they held relatively strong prior causal beliefs, potentially explaining why post-learning causal judgments remained above zero. However, the authors noted that their attempt to manipulate priors was unsuccessful (p. 5), limiting the extent to which observed judgments could be attributed to prior beliefs. Whether causal structure priors exert the predicted influence therefore remained an open empirical question. In addition, participants' judgments in that study were not compared to quantitative predictions from a computational model.

In the present studies, we successfully manipulated participants' causal structure priors using visual depictions of causal mechanism information (Experiments 1–3) and compared their causal judgments to the predictions of the SI model (Meder et al., 2014). Overall, the results matched the model predictions, strengthening the rational Bayesian causal structure learning account and its explanation of the causal illusion effect. We also replicated and generalized the effects of prior beliefs using a qualitatively different manipulation: explicit statistical information about devices (Experiment 4). Furthermore, our model fit analyses, together with the results from Experiment 4 and a supplementary study, indicated that participants primarily attended to structural causal information, while there was less sensitivity to information about the sizes of the parameters. This is in line with the view that our inferential systems prioritize structure over strength (see, e.g., Sloman, 2005).

To the best of our knowledge, our results are the first to demonstrate that apparent causal illusions under zero-contingency data can be brought to near-zero and decoupled from outcome density, particularly in conditions in which the absence of a causal link is suggested. It is also

worth noting that the rating scale used in our experiments, a unidirectional slider, has previously been found to inflate the magnitude of the outcome-density effect (Ng et al., 2024). The fact that we nonetheless observed a near-complete attenuation of the effect in the transparent low-prior conditions strengthens our findings.

The present findings also invite a reconsideration of how the causal-illusion effect is characterized in applied research. A substantial literature has studied the tendency to perceive causal connections under zero contingency across clinical and healthy populations — under labels such as illusion of control" (Alloy & Abramson, 1979) — and has found that individuals with depression tend to provide causal judgments closer to zero than healthy controls, a pattern sometimes described as "depressive realism" or "sadder but wiser." From a bias perspective, this has been interpreted as evidence that healthy individuals are subject to a cognitive distortion from which depressed individuals are paradoxically spared. Our results suggest a different reading: if non-zero causal judgments under zero contingency reflect rational Bayesian updating from non-zero priors, then the reduced judgments observed in depressed individuals may indicate that depression is associated with lower prior causal beliefs, rather than superior normative reasoning. Conversely, the elevated judgments of healthy individuals need not index irrationality. Disentangling these possibilities will require measuring participants' prior causal beliefs directly, rather than inferring their rationality from the magnitude of their post-learning judgments alone. It is worth noting, however, that the depressive realism effect itself has proven difficult to replicate: a recent well-powered, preregistered study found no evidence that symptoms of depression are associated with reduced illusory control (Dev et al., 2022). Disentangling these possibilities will require measuring participants' prior causal beliefs directly, rather than inferring their rationality from the magnitude of their post-learning judgments alone.

Our results also shed light on the limits of rational computational Bayesian models. Across experiments, we consistently observed a minority of participants who exhibited non-normative behavior: even those who held near-zero priors sometimes upward-adjusted their ratings after zero-contingency data, suggesting genuine "causal illusions" in the classic sense. Whether these individuals differ systematically in cognitive style, numeracy, or engagement with the stimuli is an important question for future work. The substantial interindividual variability observed here is not predicted by the SI model and points to the need for mixture models or individual-difference accounts that can accommodate both rational and non-rational updating strategies within the same framework.

The presence of these individual differences also highlights a broader limitation of purely computational-level accounts. While important for characterizing normative inference, they are not sufficient for a complete explanation of cognitive phenomena. In this regard, we agree with Jones and Love's (2011) critique of "Bayesian Fundamentalism." A rational Bayesian analysis can demonstrate that a particular pattern of behavior is (more or less) consistent with optimal inference, but because the psychological processes that generate the behavior are left unspecified, it does not explain, for example, why some individuals depart systematically from those predictions, nor what cognitive mechanisms give rise to such departures. Understanding the origins of causal illusions therefore requires complementing computational-level accounts with theories of individual differences, representations, and cognitive processes that determine how evidence is encoded and integrated.

Beyond their implications for Bayesian theories of causal learning, the present results speak to the question of how causal structure priors are formed in everyday cognition. Outside the laboratory, learners rarely approach causal questions as blank slates; they bring background knowledge about the entities involved, their known capacities, and the plausibility of causal connections between them. This knowledge shapes what people expect before any data are observed. Our results suggest that this prior knowledge can be potent enough to substantially alter the conclusions drawn from identical statistical evidence. Two learners observing the same data may reach different causal conclusions not because one is reasoning poorly, but because they enter the learning task with different — and potentially equally justified — prior beliefs. This has implications beyond the laboratory: apparent disagreements about causal conclusions in everyday life, science communication, or even legal reasoning may sometimes reflect differences in priors rather than differences in the interpretation of evidence.

It is also important to situate the present findings within the broader landscape of causal illusion research. We here focused on apparent causal illusions related to outcome density. This effect is only one of several related phenomena in which statistical evidence is systematically misread. Related effects include the cue-density effect, in which causal judgments are inflated by a high frequency of the *cause* rather than the effect (e.g., Perales et al., 2017; Perales & Shanks, 2007; White, 2003), and pseudo-contingencies, in which people infer a relationship between two variables based on their shared distributional skew alone, even in the absence of any actual covariation (e.g., Fiedler & Kutzner, 2017). While our results demonstrate that causal illusions resulting under zero contingencies and variations in outcome density admit a rational explanation, this should not be taken to imply that all forms of causal illusions can be explained rationally. There can be little doubt that genuinely non-normative causal beliefs exist and are widespread (Matute et al., 2015). Superstitious thinking is pervasive, even among educated adults: rituals, lucky charms, and unfounded causal attributions are documented across cultures (see Vyse, 2013). Such beliefs are difficult to attribute solely to non-zero priors or ambiguous data; they often persist in the face of abundant disconfirming evidence. The SI model and related frameworks may explain some forms of "causal illusion" as rational inference under uncertainty, but there certainly also exist genuinely non-normative causal beliefs. Understanding what distinguishes rational from non-rational causal inference, both at the computational and process levels, remains a challenge for future research.

In future studies, we aim to develop materials that allow for finer-grained manipulation of causal structure priors beyond the two or three levels tested here. Another limitation of the present work is that we held sample size constant across experiments, presenting participants with the same number of learning observations in all conditions. According to Bayesian updating models, sample size should matter (see, e.g., Griffiths & Tenenbaum, 2005, for empirical evidence): under constant zero contingency, larger samples provide stronger evidence against a causal link and should drive posterior beliefs closer to zero, whereas under constant positive contingency, larger samples should increase confidence in a causal link. This means that even a reasoner with a high prior belief in the existence of a causal link should eventually arrive at a low posterior belief if the zero-contingency data set is large. Whether the formation of apparent causal illusions is modulated by sample size in the way predicted by the SI model, and whether mechanism-based priors interact with sample size during updating, remains an open empirical question we aim to address in future work. A further goal is to investigate whether the approach generalizes beyond artifact scenarios to other causal domains, such as biological or social domains. It is an open question whether mechanism knowledge takes similar or qualitatively different forms across different domains. Finally, developing a principled formal account of how rich mechanism information, as described by the New Mechanism view, maps onto the priors of Bayesian causal learning models remains an important theoretical challenge.

CRediT authorship contribution statement

Simon Stephan: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Michael R. Waldmann:** Writing – review & editing, Funding acquisition, Conceptualization.

Ethics declaration

Written informed consent to take part in the study and to publish the article has been obtained from all participants or their legal representatives. The privacy rights of participants have been observed.

This study was performed in compliance with relevant laws, regulatory frameworks and guidelines where the research took place. This study was approved by the Ethics Committee of the Institute of Psychology at the University of Göttingen. (Approval No. 285)

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Sarah Plací and Tom Wysocki for helpful comments. We thank the team of reviewers for the constructive and helpful evaluation. Portions of this work have been presented in the Proceedings of the Cognitive Science Conference:

Stephan, S., & Waldmann, M. (2026). Manipulating Causal Priors Affects the Outcome-Density Bias. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 48. Retrieved from https://escholarship.org/uc/item/8gw386r2

Data availability

Data and analyses are publicly available in a GitHub repository: https://github.com/SimonStephan31/Priors_and_Causal-Illusions.

References

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14*, 381–405. http://dx.doi.org/10.1016/0023-9690(83)90024-3.

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310. http://dx.doi.org/10.1073/pnas.1912341117.

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General*, *108*(4), 441–485. http://dx.doi.org/10.1037/0096-3445.108.4.441.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233. http://dx.doi.org/10.1016/0001-6918(80)90046-3.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models Usinglme4. *Journal of Statistical Software*, *67*(1), http://dx.doi.org/10.18637/jss.v067.i01.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. http://dx.doi.org/10.1073/pnas.1306572110.

Blanco, F., & Matute, H. (2019). Base-rate expectations modulate the causal illusion. *PLOS ONE*, *14*(3), Article e0212615. http://dx.doi.org/10.1371/journal.pone.0212615.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(6), 1119–1140.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning*, *8*(4), 269–295. http://dx.doi.org/10.1080/13546780244000060.

Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, *12*(4), 353–378. http://dx.doi.org/10.1080/13546780500368965.

Cartwright, N. (2019). *Nature, the artful modeler. Lectures on laws, science, how nature arranges the world and how we can arrange it better*. Chicago: Open Court.

Cartwright, N., Pemberton, J., & Wieten, S. (2020). Mechanisms, laws and explanation. *European Journal for Philosophy of Science*, *10*, 25. http://dx.doi.org/10.1007/s13194-020-00284-y.

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

Chow, J. Y. L., Colagiuri, B., & Livesey, E. J. (2019). Bridging the divide between causal illusions in the laboratory and the real world: the effects of outcome density with a variable continuous outcome. *Cognitive Research: Principles and Implications*, *4*(1), 1–15. http://dx.doi.org/10.1186/s41235-018-0149-9.

Dev, A. S., Moore, D. A., Johnson, S. L., & Garrett, K. T. (2022). Sadder ≠ wiser: Depressive realism is not robust to replication. *Collabra: Psychology*, *8*(1), 38529. http://dx.doi.org/10.1525/collabra.38529.

Fiedler, K., & Kutzner, F. (2017). Pseudocontingencies. In M. R. Waldmann (Ed.), *The oxford handbook of causal reasoning* (pp. 189–200). Oxford University Press.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events.. *Psychological Review*, *128*, 936–975, https://psycnet.apa.org/doi/10.1037/rev0000281.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford: Oxford University Press.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Gong, T., Pacer, M., Griffiths, T. L., & Bramley, N. R. (2026). Rational causal induction from events in time. *Psychological Review*, *133*(3), 584–618. http://dx.doi.org/10.1037/rev0000570.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*(5), 1205–1222. http://dx.doi.org/10.1111/1467-8624.00224.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, *35*(8), 1407–1455. http://dx.doi.org/10.1111/j.1551-6709.2011.01203.x.

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

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

Illari, P. M. K., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*, 119–135.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*, 1–17.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188. http://dx.doi.org/10.1017/S0140525X10003134.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. http://dx.doi.org/10.1037/h0034747.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1–17. http://dx.doi.org/10.1017/S0140525X0004114X.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik, & L. Schulz (Eds.), *Causal learning: psychology, philosophy, and computation* (pp. 154–172). Oxford University Press, http://dx.doi.org/10.1093/acprof:oso/9780195176803.003.0011.

Le Pelley, M. E., Beesley, T., & Griffiths, O. (2017). Associative accounts of causal cognition. In M. R. Waldmann (Ed.), *The oxford handbook of causal reasoning* (pp. 13–28). Oxford: Oxford University Press, http://dx.doi.org/10.1093/oxfordhb/9780199399550.013.2.

Lober, K., & Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). *Psychological Review*, *107*(1), 195–212.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–982.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*, 1–25.

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, *6*, 888. http://dx.doi.org/10.3389/fpsyg.2015.00888.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning.. *Psychological Review*, *121*(3), 277–301.

Msetfi, R. M., Murphy, R. A., Simpson, J., & Kornbrot, D. E. (2005). Depressive realism and outcome density bias in contingency judgments: The effect of the context and intertrial interval. *Journal of Experimental Psychology: General*, *134*(1), 10–22. http://dx.doi.org/10.1037/0096-3445.134.1.10.

Musca, S. C., Vadillo, M. A., Blanco, F., & Matute, H. (2010). The role of cue information in the outcome-density effect: evidence from neural network simulations and a causal learning experiment. *Connection Science*, *22*(2), 177–192. http://dx.doi.org/10.1080/09540091003623797.

Ng, D. W., Lee, J. C., & Lovibond, P. F. (2024). Unidirectional rating scales overestimate the illusory causation phenomenon. *Quarterly Journal of Experimental Psychology*, *77*(3), 526–541. http://dx.doi.org/10.1177/17470218231175003.

Ng, D. W., Lee, J. C., & Lovibond, P. F. (2026). The role of prior beliefs in causal illusions. *Cognition*, *266*, Article 106290. http://dx.doi.org/10.1016/j.cognition.2025.106290.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, MA: Cambridge University Press.

Perales, J. C., Catena, A., Cándido, A., & Maldonado, A. (2017). Rules of causal judgment: Mapping statistical information onto causal beliefs. In M. R. Waldmann (Ed.), *The oxford handbook of causal reasoning* (pp. 29–51). Oxford: Oxford University Press, http://dx.doi.org/10.1093/oxfordhb/9780199399550.013.3.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review*, *14*(4), 577–596. http://dx.doi.org/10.3758/BF03196807.

Rottman, B. M. (2017). The acquisition and use of causal structure knowledge. In M. R. Waldmann (Ed.), *The oxford handbook of causal reasoning* (pp. 85–114). New York: Oxford University Press.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511623271.

Shanks, D. R., López, F. J., Darby, R. J., & Dickinson, A. (1996). Distinguishing associative and probabilistic contrast theories of human contingency judgment. In *The psychology of learning and motivation*: *vol. 34*, (pp. 265–312). San Diego, CA: Academic Press, http://dx.doi.org/10.1016/S0079-7421(08)60563-0.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2022). Afex: Analysis of factorial experiments. URL https://CRAN.R-project.org/package=afex. (R package version 1.1-1).

Sloman, S. A. (2005). *Causal models: How we think about the world and its alternatives*. Oxford: Oxford University Press.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*(3), 303–333. http://dx.doi.org/10.1207/s15516709cog2803_1.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York, NY: Springer-Verlag.

Spohn, W. (2001). Bayesian nets are all there is to causal dependence. In M. C. Galavotti, P. Suppes, & D. Costantini (Eds.), *Stochastic causality* (pp. 157–172). Chicago: University of Chicago Press.

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation – a computational model. *Cognitive Science*, *44*(7), Article e12871. http://dx.doi.org/10.1111/cogs.12871.

Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, *10*(1), 242–257.

Ullman, T. D., Spelke, E. S., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665. http://dx.doi.org/10.1016/j.tics.2017.05.012.

Vadillo, M. A., Musca, S. C., Blanco, F., & Matute, H. (2011). Contrasting cue-density effects in causal and prediction judgments. *Psychonomic Bulletin & Review*, *18*(1), 110–115. http://dx.doi.org/10.3758/s13423-010-0032-2.

Vyse, S. A. (2013). *Believing in magic: The psychology of superstition - updated edition*. New York: Oxford University Press.

White, P. A. (2003). Making causal judgments from the proportion of confirming instances: the pCI rule. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(4), 710–727.