Manipulating Causal Priors Affects the Outcome-Density Bias

Simon Stephan (simon.stephan@psych.uni-goettingen.de) & Michael R. Waldmann
Department of Psychology, University of Göttingen

Abstract

Non-contingent learning data can lead to the (seemingly) illusory perception of causality if the potential cause and effect often co-occur – an effect known as "outcome-density bias." Bayesian models of causal induction explain this effect as the result of a *rational* learning process in which the data do not fully override non-zero causal priors. Convincing evidence for this rational explanation requires the demonstration of an experimentally manipulated effect of causal priors, which has been lacking. We successfully manipulated participants' ($N = 300$) causal priors through visually conveyed causal mechanism information. This manipulation influenced the outcome-density effect in the predicted way: in a condition inducing low causal priors, the effect disappeared. The results demonstrate the effectiveness of visual mechanism information in manipulating priors and strengthen the Bayesian view of causal induction.

Keywords: outcome-density bias, causal induction, Bayesian learning, computational modeling, causal support, mechanisms

Introduction

Causal learning (often) requires the assessment of the contingency between potential cause C and effect E (see Le Pelley et al., 2017; Perales et al., 2017; Rottman, 2017, for overviews). Consider a toddler playing with a wooden block, a rubber ball, and a handheld flashlight. The child first bangs the block and then rolls the ball, but neither action produces light. However, when the child presses a specific button on the flashlight, it illuminates immediately. By repeatedly pressing the button and the observation of the following positive contingency, the child identifies the causal link. When later asked to "show me the light," the child ignores the other toys and promptly presses the flashlight button. The child has inferred the generative cause through contingency assessment.

This (anecdotal) scenario is only one example illustrating that causal induction based on contingency information often succeeds. Other, seemingly less successful cases resulting in so-called "causal illusions" have also received the attention of researchers, however. One of these cases is the so called "outcome-density bias" (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). It refers to the phenomenon that under constant levels of contingency (ΔP) between a putative cause C and an effect E, judgments of causality tend to increase with the marginal probability of the effect, $P(E)$. Crucially, this effect persists even when there is *no* empirical contingency between C and E, that is, when $\Delta P = 0$. Therefore the term "causal illusion."

What explains this apparent outcome-density bias, and which factors other than $P(e^+ \mid c^-)$ influence it? Accounts that characterize this effect as an irrational bias typically explain it in terms of differential weighting of trial types. A prominent weighting scheme that has been proposed (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, the outcome-density bias arises because reasoners place disproportionate weight on "confirming" observations – especially trials in which C and E co-occur ($N_i(e^+ \mid c^+)$ or a trials) – and comparatively little weight on disconfirming observations (e.g., $N_i(e^+ \mid c^-)$ or c trials).

Griffiths and Tenenbaum (2005) argued that the effect can be explained by Bayesian causal structure learning, formalized in their causal support model (Griffiths & Tenenbaum, 2005) or in the later structure induction (SI) model (Meder et al., 2014). According to this view, the effect does not reflect an irrational bias but instead arises *normatively* when a reasoner assigns a non-zero *prior probability* to the existence of a causal link between C and E, and this prior is not fully overridden by the learning data.[1] This leads to the hypothesis that the magnitude of the outcome-density effect should depend on the degree to which a reasoner believes that C causes E prior to seeing data.

A limitation of previous work on Bayesian models of the outcome density effect is that causal structure priors have neither been directly measured nor successfully experimentally manipulated (but see Ng et al., 2026). It remains an open empirical question whether differences in prior causal beliefs influence the outcome-density effect in the predicted way. We here address this question and empirically show that it does. Moreover, we demonstrate that the effect can be brought to disappear if participants hold close-to-zero causal priors. Our manipulation of participants' causal priors targeted the causal mechanism knowledge they brought to bear on the causal learning task. Our findings corroborate the Bayesian learning account of causal induction and its explanation of the outcome-density effect. We next summarize the structure-induction view on the outcome-density effect

[1] We henceforth adopt the neutral formulation "outcome-density effect."

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

and then describe our visual mechanism-centered approach to manipulating causal priors.

The outcome-density effect through the lens of Bayesian causal structure learning

Many theories of causal induction model causal learning as "causal strength learning" (Buehner et al., 2003; Cheng, 1997; Jenkins & Ward, 1965; Lober & Shanks, 2000; Shanks et al., 1996) based on point estimates directly determined by the observed data. These theories fail to predict the outcome-density effect (but see Cheng, 1997, for a model that explains this effect when $\Delta P > 0$).

By contrast, it has been argued that causal induction is often better characterized as "causal structure learning" (Griffiths & Tenenbaum, 2005, 2009). Causal structure learning addresses *whether* a causal link between two event types exists, whereas causal strength learning estimates the magnitude of such a link, conditional on its existence.

A core assumption implemented in computational models of causal structure learning is that learning follows the principles of Bayesian updating. Here, we use the SI model (Meder et al., 2014; Stephan & Waldmann, 2018) originally inspired by Griffiths and Tenenbaum's *causal support* model, to illustrate how Bayesian causal structure learning accounts for the outcome-density effect, and to demonstrate the predicted influence of causal structure priors.[2]

According to the model, a causal learner uses empirical data as evidence to update prior probabilities over alternative causal hypothesis about the data. These hypotheses are formally represented as causal Bayes nets (Glymour, 2001; Pearl, 2000; Spirtes et al., 1993; Spohn, 2001). Under causal structure S_1, C is a cause of E, in addition to unobserved independent background causes A ($S_1 : C \rightarrow E \leftarrow A$). Accordingly, some of the co-occurrences of C and E observed in a contingency table are explained by the causal influence of C, other cases of co-occurrence are a coincidence (caused by A). Under the competing causal structure, S_0, no causal link between C and E is assumed; E is influenced solely by the background causes A ($S_0 : C \quad E \leftarrow A$). All observed co-occurrences between C and E are treated as coincidental under S_0. Learning is assumed to update the 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). It suffices here to note that Eq. 1 entails that the posterior probability of a causal link between C and E ($P[S_1 \mid D]$) is proportional to the product of the likelihood of the data under S_1 ($P[D \mid S_1]$) *and* its prior probability ($P[S_1]$):

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

The default assumption of the SI model in the literature is that a reasoner approaches the learning task with an uninformative structure prior, that is, they regard S_0 and S_1 as equally plausible: $P(S_1) = P(S_0)$ (but see Lu et al., 2008). Because S_1 and S_0 are mutually exclusive and exhaustive causal hypotheses about the observed data, it follows that $P(S_1) = 1 - P(S_0)$ and, conversely, that $P(S_0) = 1 - P(S_1)$. A uniform prior therefore implies $P(S_1) = P(S_0) = 0.5$.

Assuming that an uninformative prior captures participants' assumptions in previous studies, the SI model predicts the outcome-density effect. This is illustrated in Fig. 1, which shows for different zero-contingency data sets how the posterior belief in structure S_1, ($P[S_1 \mid D]$, y-axis), changes with the prior belief ($P[S_1]$, x-axis). 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. Fig. 1 shows that this is also the case for S_1 priors other than 0.5, although the influence of $P(e^+ \mid c^-)$ becomes weaker the more the S_1 priors approach the extremes.

Under ceiling conditions, $P(e^+|c^+) = P(e^+|c^-) = 4/4$ in Fig. 1, the model even predicts a mild increase from prior to posterior for any prior value between 0 and 1.0. As shown in Fig. 1 (see also Fig. 2), this increase is predicted not only for uninformative structure priors but for all priors greater than zero and less than one. It is predicted because ceiling-effect data leave the causal strength of the target cause C underdetermined (cf. Cheng, 1997). The observed effects could be generated exclusively by background causes A, or C could additionally contribute but be masked by A's influence. Because Bayesian model averaging evaluates causal structures by integrating over parameter uncertainty, structure S_1 retains a broader range of parameter values consistent with the observed data: no value of the causal strength parameter w_c can be ruled out. Consequently, rational Bayesian updating can lead to an increase of the posterior probability of S_1, even when the empirical contingency is zero.

The magnitude of this effect under ceiling conditions depends on the assumed priors over the model parameters and on the assumed functional integration of C and A under S_1. The SI model and the causal support model assume flat priors over the structures' strength parameters (see Lu et al., 2008, who suggested a "strong and sparse" prior rather than flat strength priors), and a noisy-OR integration function. To the best of our knowledge, this effect has not previously been discussed theoretically or demonstrated empirically.

The plausibility of an uninformative structure prior depends on the knowledge (or lack thereof) that a reasoner brings to the learning context (cf. Griffiths & Tenenbaum, 2009). In typical

[2] Both models make identical assumptions, but the SI model yields the posterior probability of a causal link between C and E, whereas the causal support model computes the log likelihood ratio between causal structures with and without such a link. The support measure can be converted into a posterior probability through a logistic transformation (cf. Griffiths & Tenenbaum, 2009).

Figure 1: Relation between the prior probability, $P(S_1)$, and the posterior probability, $P(S_1|D)$, of S_1 across different zero-contingencies, computed by the SI model.

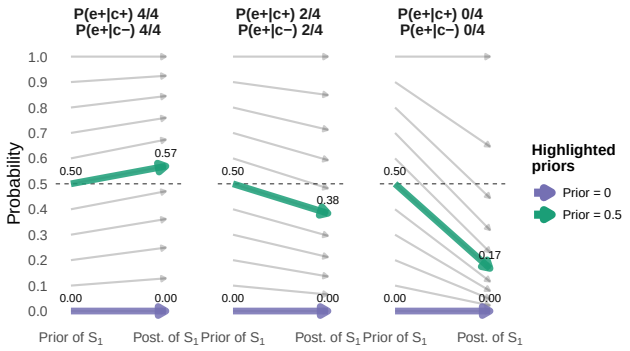

Figure 2: $P(S_1)$ and $P(S_1|D)$ pairs for different S_1 priors. Highlighted priors (of $P(S_1) = 0.5$ and $P(S_1) = 0$) are those supposed to be the ones assumed by participants in the experiment's different device mechanism conditions.

cover stories, structure priors close to 0.5 appear plausible. By contrast, in situations in which domain-specific knowledge is available (cf. Griffiths & Tenenbaum, 2009), more informative priors seem more appropriate for generating model predictions. Importantly, as shown in Fig. 1, any prior belief in S_1 greater than zero yields posteriors greater than zero, with larger outcome-density effects for stronger S_1 priors. Crucially, the figure also shows that a near-zero posterior belief in S_1 (i.e., the disappearance of the outcome-density effect) can only arise if a reasoner starts with a near-zero causal prior. Our objective was to create an experimental condition that encourages such a prior.

An alternative graph showing the predictions made by the SI model for different structure priors is shown in Fig. 2. The structure priors we sought to induce in the different conditions of our experiment, together with resulting posteriors, are highlighted in green (uninformative prior) and violet (zero prior), respectively.

Methods of manipulating priors

Verbal cover stories

How can a learner's causal priors be successfully manipulated? 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). Thus, one possible approach to manipulate priors could be to vary the domain of the cover story. This approach was recently taken 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 induce different causal priors, the approach comes with a lack of experimental control. Different cover stories may differ along dimensions other than the intended ones (causal priors in this case).[3] Another problem is that cover stories about potential cause-effect relations fail to fully resolve the ambiguity whether an observed pairing of a potential cause and effect event is causal or coincidental. Causal pairings boost the likelihood of the causal structure model (S1), whereas non-causal pairings boost the likelihood of the non-causal structure model (S0).

Visual mechanism information

We here sought to manipulate causal priors by varying *visual causal mechanism* information, while holding other aspects of the scenario constant. Specifically, we sought to capitalize on reasoners' ability to readily interpret visually conveyed causal mechanisms if those are presented in a tractable manner. What we take to be a "tractable" visual representation of a causal mechanism is inspired by the so-called New Mechanism view (Cartwright, 2019; Cartwright et al., 2020; Glennan, 2017; Illari & Williamson, 2012; Machamer et al., 2000), which gives the following characterization of a mechanism: "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 therefore be involved in the production of change (Glennan, 2017) (i.e., causal processes). Fig. 3 shows our visual stimuli. Although the displays show static devices, the processes they can give rise to can be easily simulated based on knowledge about the entities, their capacities, and their configuration (see also Allen et al., 2020; Battaglia et al., 2013; Ullman et al., 2017, for another physics simulation account).

The fictitious experimental scenario we developed involves an unfamiliar kind of mechanical device allegedly used by developmental psychologists to study causal learning. Illustrations are shown in Fig. 3, where Fig. 3A depicts the external appearance and Fig. 3B–D show the mechanisms of different versions of the device.

[3] Highly parallelized cover stories maximizing control might elicit similar priors, whereas highly different cover stories introduce the risk of confounding.

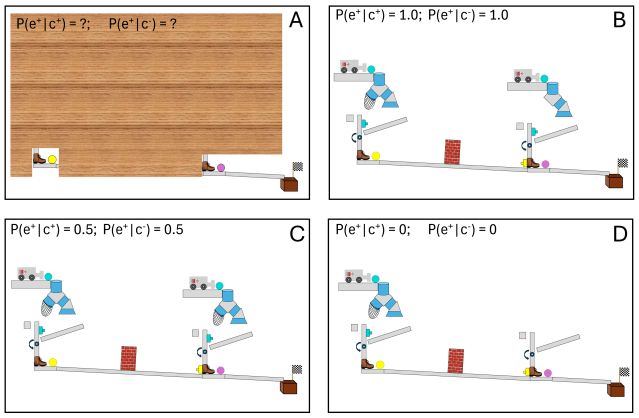

Figure 3: Experimental stimuli. A: opaque mechanism. B-D: transparent mechanisms generating different contingencies. The included probability information was not shown to participants in the experiment.

The scenario description explicitly mentions that there are two types of the device: half of the existing devices are "causal" ones (instantiating S_1), in which rolling of the yellow ball can cause rolling of the pink ball, and the other half are "non-causal" ones (instantiating S_0), in which rolling of the yellow ball cannot cause rolling of the pink ball. The instruction also points out that the pink ball may roll even if the yellow ball does not, due to an alternative launching mechanism inside the device. Moreover, it is emphasized that all devices look the same from the outside.

The different mechanism illustrations shown in Fig. 3B–D were intended to convey that a vehicle above the yellow ball pushes a light blue ball into a forking tube system. The light blue ball can take one of two possible paths, a left one leading into a net, or a right one that leads onto a slope leading to a lever. A light blue ball rolling down that slope triggers the lever, which launches the yellow ball. This triggering component was the same in all devices. It was supposed to convey the impression of a 50% chance of the yellow ball rolling (C).

A launched yellow ball rolls down a slope towards the pink ball (E). Unless disturbed, it has the capacity to push a second lever connected the pink ball. However, a wall is positioned on the path blocking the yellow ball so that it cannot reach the pink ball. The pink ball still can roll when a second triggering component is active, similar to the one of the yellow ball. This part differs between the different device versions, to convey different probabilities of the effect occurring: 1.0 in Fig. 3B, 0.5 in Fig. 3C, and 0 in Fig. 3D. Apart from these visual depictions, no additional information about how the shown mechanisms work was provided.

The stimuli revealing the mechanisms are supposed to saliently convey that, due to the blocking of the path of the yellow ball by the wall, there is no plausible causal link between the yellow and pink ball. These stimuli also show why the pink ball sometimes moves due to a cause independent of the yellow ball. Thus, all observed co-occurrences during the learning phase between the movements of the yellow and the pink ball should be interpreted as coincidences. This setup suggests a strong prior for S_0. If this is impression is indeed conveyed by the materials, and if structure priors directly influence the outcome-density effect, then the effect should disappear in participants who see the causal mechanism stimuli. By contrast, participants who only see the device from the outside (cf. Fig. 3A) but not the device mechanisms should tend towards uninformative causal priors, leading to the typical outcome-density effect.

Experiment

Our experiment was designed to test these predictions. The experimental materials and analysis files are available in a GitHub repository at https://github.com/SimonStephan31/Manipulating-Prior-Causal-and-Outcome-Density-Bias.

Methods

Participants Four hundred and four participants were recruited via www.prolific.co. 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[4]). One participant was excluded for failing to confirm participation via a PC or laptop. An additional 101 participants were excluded because they did not meet a prespecified learning-phase performance criterion (see Procedure). The final analysis sample consisted of $N = 300$ participants. Data collection was terminated once all experimental conditions included $n = 50$ participants, after exclusions.

Design, Materials, and Procedure The study employed a 2 (device mechanism: opaque vs. transparent; 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 a verbal description of the scenario about developmental psychologists who built an unfamiliar kind of device allowing them to study causal learning in children. Then participants were told that the devices can generate movements of a yellow and a pink ball. Together with this information, participants were shown a picture showing the device with the mechanism covered (Fig. 3A). They also learned that in one type of these devices, the rolling of the yellow ball can cause the pink ball to roll (instantiating S_1). In another type – which, unbeknownst to participants, was the only one actually shown during the experiment – the rolling of the yellow ball cannot cause the pink ball to roll (instantiating S_0).

[4]We established this criterion to ensure that participants understand the instructions.

They also learned that, in the "causal" devices, the probability with which rollings of the yellow lead to rollings of the pink ball may vary between devices. Additionally, it was mentioned that a device may contain an alternative triggering mechanism which, independent of the yellow ball, can also generate rollings of the pink ball.

Participants were then informed that they would study one specific device and that their task was to determine whether a rolling of the yellow ball can cause a rolling of the pink ball in this device. In the transparent mechanism conditions, participants were additionally shown an image revealing the mechanism of the specific target device. Depending on the contingency condition, the mechanism picture they were shown was one of the stimuli depicted in Fig. 3B–D. These pictures were 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 either showed the opaque device again or the picture revealing the device's mechanism. The test query, presented below the picture, read: "How confident 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 using a 100-point continuous slider with a one-point resolution (endpoints: "Rolling of the yellow ball CANNOT cause rolling of the pink ball" and "Rolling of the yellow ball CAN cause rolling of the pink ball," midpoint: "Uncertain").

Participants were next informed about the learning phase. They learned that they would be shown eight static pictures in which the device was activated, which would be presented trial-by-trial. Participants in the transparent mechanism conditions were informed that the device's inside would not be visible anymore during this phase. To support learning and to establish a performance criterion, participants indicated on each trial which of the two balls rolled, using radio buttons displayed below the stimulus. Participants who made errors on more than one of the eight trials were excluded from the analyses.

After the learning phase, participants responded to the same causal structure query again, to assess their structure posterior.

Results and Discussion

Participants' causal structure ratings are shown in Fig. 4. As predicted, the experimental manipulation had the intended effect: Participants in the opaque conditions gave relatively high posterior causal ratings, exhibiting a clear outcome-density effect. This pattern is consistent with participants holding structure priors close to 0.5. By contrast, participants in the transparent visual mechanism conditions gave posterior ratings close to zero and showed no outcome-density effect. According to the theory, the lack of this effect is a consequence of the fact that participants entered the learning phase with a low S_1 causal structure prior. By contrast, participants in the opaque conditions were more strongly influenced by $P(e^+ \mid c^-)$ than those in the transparent conditions, with the

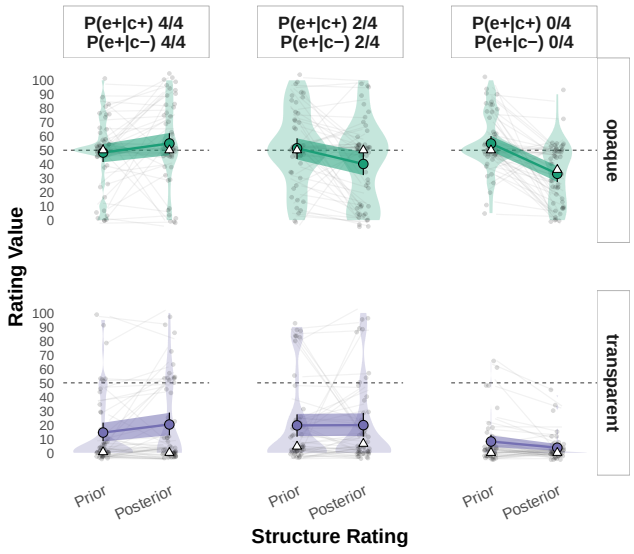

Figure 4: Participants' prior and posterior causal structure ratings. Colored circles are means and error bars their 95% CIs. Triangles are medians. Jittered dots and lines are individual ratings, and violin plots their distribution.

outcome-density effect diminishing as $P(e^+ \mid c^-)$ decreased. Overall, participants' causal ratings matched the predictions of the SI model (cf. Fig. 2).

These observations were corroborated by a mixed ANOVA with causal structure rating (prior vs. posterior) as a within-subjects factor and contingency and device mechanism knowledge as between-subjects factors, using the `afex` package in R (Singmann et al., 2022). We report only theoretically relevant effects. There was a significant main effect of device mechanism knowledge, $F(1, 294) = 171.57$, $p < .001$, $\eta_p^2 = .369$, reflecting higher causal ratings in the opaque than in the transparent conditions. There was also a significant main effect of causal rating type, $F(1, 294) = 8.67$, $p < .001$, $\eta_p^2 = .029$, indicating lower posterior than prior ratings overall. Crucially, this effect depended on device mechanism knowledge, as shown by the significant interaction between causal rating type and device mechanism, $F(1, 294) = 10.88$, $p < .001$, $\eta_p^2 = .036$: the reduction from prior to posterior ratings was stronger in the opaque than in the transparent conditions.

In addition, there was a significant interaction between causal rating type and contingency, $F(2, 294) = 15.77$, $p < .001$, $\eta_p^2 = .097$, reflecting differential updating across contingency conditions. Importantly, this interaction also interacted with device mechanism knowledge, as indicated by a significant three-way interaction, $F(2, 294) = 3.54$, $p = .03$, $\eta_p^2 = .023$. This pattern matches the predictions of the SI model (cf. Fig. 2 and Fig. 4).

We additionally examined participants' causal ratings at the individual level by classifying the changes between prior and posterior judgments. The results are shown in Fig. 5. In the opaque-mechanism conditions, the proportion of participants whose posterior ratings were lower than their prior ratings

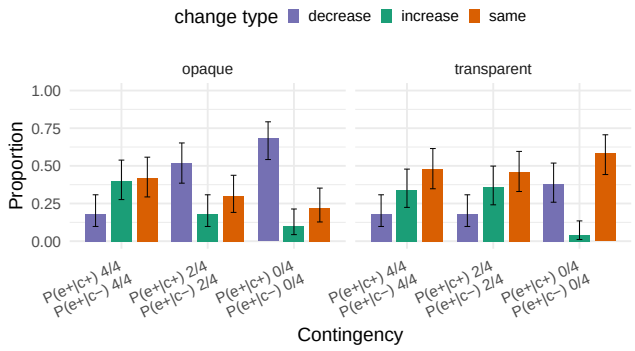

Figure 5: Proportions of observed rating-change types. Error bars show 95% Wilson CIs.

(the "decrease" cluster) increased as $P(e^+ \mid c^-)$ decreased. This pattern is consistent with the aggregate analyses, indicating that the more diagnostic the data were for a non-causal structure (i.e., the larger $P(D \mid S_0) > P(D \mid S_1)$), the more participants downward-adjusted their initial beliefs.

By contrast, in the transparent-mechanism conditions, the largest cluster across all contingency conditions consisted of participants who gave identical prior and posterior ratings. Accordingly, the size of the "decrease" cluster remained approximately constant across contingency conditions in these conditions.

An additional but noteworthy finding is that a subset of participants *increased* their causal ratings from prior to posterior in the ceiling-effect contingency condition ((4/4)(4/4); cf. Fig. 4 and Fig. 5), a pattern that is predicted by the SI model, as shown in Figure 1 and 2 (see also Griffiths & Tenenbaum, 2009, p. 676). This effect provides additional evidence for the Bayesian nature of human causal structure induction and for the assumptions underlying the SI model.

Although the results were largely in line with the Bayesian learning explanation of the outcome-density effect, the individual ratings shown in Fig. 4 indicate that the behavior of at least a number of participants deviated from the predictions. For example, there were some participants who increased their ratings from prior to posterior, even in the non-ceiling conditions where such an effect is not predicted by the SI model. This means that a minority of participants may have formed "irrational" causal impressions. These residual genuine outcome-density biases may have resulted from biased weighting of the learning trials (cf. Perales et al., 2017), or a lack of understanding of the mechanism.

General Discussion

When a potential cause and effect are uncorrelated in the learning data, a high frequency of the effect can nonetheless lead to a (seemingly) illusory perception of causality—a phenomenon known as the "outcome-density bias." This effect has remained a topic of debate and ongoing investigation (for recent studies, see, e.g., Blanco & Matute, 2019; Chow et al., 2024; Ng et al., 2026). Most studies have approached the outcome-density effect from a bias perspective (but see Ng et al., 2026), characterizing it as a "causal illusion" or as "incorrectly reporting a causal relationship where *there is none*" (e.g., Blanco & Matute, 2019, p. 1).

The prominence of this bias-focused perspective ignores the fact that the outcome-density effect can be given a rational explanation (but see Ng et al., 2026). From a Bayesian causal structure learning perspective (Griffiths & Tenenbaum, 2005, 2009), the effect does not reflect an illusory causal belief, but rather the recognition that the observed data do not provide sufficient evidence against the existence of a causal relationship – assuming such a relationship is a priori plausible.

One reason why this rational explanation has received relatively little attention may be that it has rarely been the focus of empirical investigation. An exception is a recent study by Ng et al. (2026), who assessed participants' causal priors before presenting learning data and found that participants held relatively strong prior causal beliefs, potentially explaining why post-learning causal judgments remained greater than zero. However, the authors noted that their manipulation of causal priors was unsuccessful, limiting the extent to which final causal judgments could be attributed to prior beliefs (p. 5). As a result, whether causal structure priors exert the predicted influence remained an open empirical question. In addition, participants' judgments were not compared to quantitative predictions derived from a computational model.

We successfully manipulated participants' causal structure priors using visual depictions of causal mechanism information, and we compared their causal judgments to the predictions of the SI model (Meder et al., 2014). The observed results matched the model predictions, thereby strengthening the rational Bayesian structure learning account of the outcome-density effect. Moreover, to the best of our knowledge, we demonstrate for the first time that the outcome-density effect can be nearly eliminated. This was achieved by providing participants with visual mechanistic information prior to learning that saliently suggested the absence of a causal link.

In future work, we plan to examine a broader range of contingency data sets, and develop materials showing mechanisms that allow for the manipulation of more than two levels of causal structure priors. We also aim to replicate the findings with causal strength rather than a causal structure test queries, to show that the Bayesian explanation of the outcome-density effect is not restricted to causal structure inferences.

Acknowledgments

We thank Sarah Placì for helpful comments. This work was supported by a Reinhart Koselleck project (WA 621/25-1), funded by the Deutsche Forschungsgemeinschaft (DFG).

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. https://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. https://doi.org/10.1073/pnas.1912341117

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. https://doi.org/10.1073/pnas.1306572110

Blanco, F., & Matute, H. (2019). Base-rate expectations modulate the causal illusion. *PLOS ONE*, *14*(3), e0212615. https://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.

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

Cartwright, N., Pemberton, J., & Wieten, S. (2020). Mechanisms, laws and explanation. *European Journal for Philosophy of Science*, *10*, 25. https://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. https://doi.org/10.1186/s41235-018-0149-9

Chow, J. Y. L., Goldwater, M. B., Colagiuri, B., & Livesey, E. J. (2024). Instruction on the scientific method provides (some) protection against illusions of causality. *Open Mind*, *8*, 639–665. https://doi.org/10.1162/opmi_a_00141

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

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

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.

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 University Press. https://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.

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

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. https://doi.org/10.1080/09540091003623797

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

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. 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 University Press. https://doi.org/10.1093/oxfordhb/9780199399550.013.3

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 University Press. https://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* (pp. 265–312, Vol. 34). Academic Press. https://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* [R package version 1.1-1]. https://CRAN.R-project.org/package=afex

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., & 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. https://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. https://doi.org/10.3758/s13423-010-0032-2

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.