Constructing causal stories: How mental models shape memory, prediction, and generalization

Steven Shin[1], Chuqi Hu[2], Paul S. Muhle-Karbe[3], and Tobias Gerstenberg[*2]

[1]Perelman School of Medicine, University of Pennsylvania
[2]Department of Psychology, Stanford University
[3]School of Psychology, University of Birmingham

Abstract

How do people learn to predict what happens next? On one account, people do so by building mental models that mirror aspects of the causal structure of the world. Accordingly, people tell a story of how the data was generated, focusing on goal-relevant information. On another account, people make predictions by learning simple mappings from relevant features of the situation to the outcome. Here, we provide evidence for the causal account. Across three experiments and two paradigms, we find that people misremember what happened, predict incorrectly what will happen, and generalize to novel situations in a way that's consistent with the causal account and inconsistent with a feature-based alternative. People spontaneously construct causal models that compress experience to privilege causally relevant information. These models organize how we remember the past, predict the future, and generalize to novel situations.

Keywords: causality; abstraction; learning; representation; mental model; intuitive physics.

[*]Corresponding author: Tobias Gerstenberg https://orcid.org/0000-0002-9162-0779 (gerstenberg@stanford.edu)

Introduction

People learn about the world by building internal mental models (Chater & Oaksford, 2013; Craik, 1943; Gerstenberg & Tenenbaum, 2017). They use these models to make predictions about the future and inferences about the past (e.g. Goodman, Tenenbaum, & Gerstenberg, 2015; Johnson-Laird, 1983; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). For example, a detective at a crime scene reconstructs what happened by mentally simulating different scenarios. Given a bullet in the wall, they can infer the angle from which it was shot. One prominent view is that people's mental model of the *physical world* may be similar to the kinds of physics engines that power realistic animations in modern computer games (Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Kubricht, Holyoak, & Lu, 2017; Smith et al., 2024; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). For a physics engine to simulate what will happen, it needs to know each object's exact location, mass, shape, friction, etc. However, people's mental models are unlikely to represent the physical world at this level of detail (Bass, Smith, Bonawitz, & Ullman, 2022; Bigelow, McCoy, & Ullman, 2023; Davis & Marcus, 2016). People have limited cognitive resources, so they cannot represent everything (Lieder & Griffiths, 2020; Sosa, Gershman, & Ullman, 2025). How do they figure out what to represent? How do they construct mental models of the world at the right level of abstraction for the task at hand?

To shed light on this question, we combine insights from work on causal abstraction (Beckers & Halpern, 2019; Chalupka, Eberhardt, & Perona, 2017) with that on goal-dependent representations in attention and planning (Ho et al., 2022; Maruff, Danckert, Camplin, & Currie, 1999). We develop a novel approach for studying causal abstraction and test it in three experiments. In Experiment 1, participants perform a simple causal learning task and, when asked in a surprise task what they just saw, make systematic memory errors that are consistent with having learned goal-dependent abstractions (see Figure 1a). Experiment 2 features a physical prediction task. Again, we find systematic errors in a surprise task that are consistent with the idea that participants learned a goal-dependent causal model of the task (see Figure 1b). While consistent with the idea that people learn abstract causal models, the results of Experiment 1 and 2 can also be explained by a feature-based model that assumes people form simple associations between predictive features and outcomes (rather than rich causal models). In Experiment 3, we show that people generalize what they've learned in a way that's consistent with the causal abstraction account and cannot be captured by the feature-based model. We conclude by discussing implications of our work as well as future research directions.

Abstract causal models

Abstract schemas and scripts organize our actions and structure our memory of what happened (Kelley, 1972, 1973; Schank & Abelson, 1977). Often, these schemas and scripts have a distinctly *causal* structure, representing not only associations between events, but also what causes what. Causal models are useful because they support reasoning about the consequences of acting on the world (Sloman, 2005; Waldmann & Hagmayer, 2005), and because they allow us to flexibly apply our knowledge in new situations (Bareinboim & Pearl, 2016; Lake, Ullman, Tenenbaum, & Gershman, 2017). Causal models can be formulated

at different levels of abstraction (Griffiths & Tenenbaum, 2009; Iwasaki & Simon, 1994). While we ultimately have to take actions in continuous space and time, we often don't think about the world that way. Instead, our mental models abstract away the lower-level details (e.g. Beckers, Eberhardt, & Halpern, 2020; Beckers & Halpern, 2019; Beller & Gerstenberg, 2025; Chalupka et al., 2017; Gerstenberg et al., 2021; Strelnikoff, Jammalamadaka, & Lu, 2022).

The question of how to choose the right level of abstraction has been studied extensively in both philosophy and cognitive science. One proposal from philosophy suggests that the variables in a causal model should be 'proportional' to one another (Woodward, 2021; Yablo, 1997). Cause and effect variables should be specified at a level of detail that matches. For example, for representing a light switch and a light bulb, binary variables work well. To represent a dimmer and a dimmable light bulb, continuous variables work well. What doesn't work well, is a continuous variable for a switch (the cause), when the bulb can only ever be on or off (the effect).

People are good at learning abstract causal models that capture the essential structure of a system while ignoring irrelevant details (Wellen & Danks, 2014). For example, adults Lu, Rojas, Beckers, and Yuille (2016) and children (Lucas, Bridgers, Griffiths, & Gopnik, 2014) can learn whether causes combine conjunctively or disjunctively to bring about effects, and generalize that knowledge to new situations. People don't just learn specific cause-effect relationships. Instead, they build abstract models that capture the underlying causal structure of the system. Although we know that people can learn abstract causal models, we know less about how they choose what information to abstract. How do people learn what features are critical for their goals?

Goal-dependent mental representations

Goals shape mental representations (Muhle-Karbe et al., 2023). They constrain what we attend to (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017; Maruff et al., 1999; Niv et al., 2015), what we learn (Flesch, Juechems, Dumbalska, Saxe, & Summerfield, 2022; Kaplan, Schuck, & Doeller, 2017; Lu et al., 2016; Mack, Love, & Preston, 2016), what we remember (Ho et al., 2022), and what we believe (Gershman, 2018; Kunda, 1990). Because people's cognitive resources are limited, they need to allocate them efficiently (Bates, Lerch, Sims, & Jacobs, 2019; Brady & Tenenbaum, 2013). For example, when people plan how to navigate a maze, they only represent what matters at a fine level of detail and represent the rest more coarsely (Ho et al., 2022). When asked to recall where an obstacle was located, they remember it well when it affected their planned route, but less well when it didn't (Ho et al., 2022). In general, there is value in abstraction (Ho, Abel, Griffiths, & Littman, 2019). Good abstractions help us learn faster and transfer that knowledge to new tasks (see also Gentner & Markman, 1997; Ho et al., 2022; Holyoak, Lee, & Lu, 2010). However, there are downsides, too. It's possible that information that we abstracted away becomes relevant when goals change.

Goal-dependent abstract causal models

Work on abstraction in causal models has mostly been theoretical and not connected directly with the way humans think (Beckers & Halpern, 2019; Chalupka et al., 2017; Zen-

naro, Turrini, & Damoulas, 2022). Work on goal-dependent representations has focused on visual attention (Maruff et al., 1999) and navigation (Ho et al., 2022). Recently, Kinney and Lombrozo (2024a) combined insights from both strands of research to develop an account of how people build compressed causal models of the world. Accordingly, any representation of a causal system has to trade off compression and informativeness (a map has to be smaller than the area it represents). People generally prefer more compressed representations: simple causal models with few variables that have a low level of granularity (Pacer & Lombrozo, 2017, see also). However, simpler representations encode less information. How should one balance the trade-off between compression and informativeness? Kinney and Lombrozo's (2024a) propose that this depends on the problem the agent faces. Compression is good as long as it doesn't lose information that's important for making good decisions (see also Williams, 1991). They quantify how much information is lost by moving from a less to a more compressed causal model by defining a causal variant of mutual information (Shannon, 1948). Intuitively, causal mutual information captures how much information an intervention in a candidate cause reveals about a candidate effect. For example, if the causal mutual information is high, then one can tell exactly what would happen to an effect if one intervened in the cause. If it's low, then a lot of uncertainty remains.

Their account unifies two features of good causal models: proportionality (which we briefly mentioned above) and stability. A causal variable is *proportional* with respect to an effect variable when it would gain little to no information if it was replaced with a more fine-grained alternative. If a light bulb can only ever be on or off then a light switch would be a proportional cause. The relationship between two variables is *stable* when the causal mutual information between cause and effect doesn't depend (much) on other background variables (Lombrozo, 2010; Woodward, 2006). Conversely, unstable causal relationships are strongly moderated by other factors (Vasilyeva, Blanchard, & Lombrozo, 2018). In several experiments, Kinney and Lombrozo (2024a) show that participants prefer simpler causal explanations when these explanations don't remove information that's relevant for the decision they have to make. In related work, they showed that people can infer what a person values from how detailed a causal explanation they give (Kinney & Lombrozo, 2024b).

Our approach

We investigate whether people learn and use goal-dependent abstract causal models. In the 'blicket paradigm' (Figure 1a), participants need to learn which objects turn on the "blicket" detector (see, e.g. Gopnik et al., 2004; Sobel & Kirkham, 2006). In the 'ramp paradigm' (Figure 1b), participants need to learn which cubes on what ramps cross a finish line (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015).

In both paradigms, there are two pairs of features that (potentially) affect the outcome. In the blicket paradigm, each object has one of two shapes, and one of two colors. In the example in Figure 1a, the square objects are blickets and the round ones aren't (and the color doesn't matter). If a learner has to predict whether a cube will activate the blicket detector, they may learn to discard some of the information (the object's color) and only keep what they need (the object's shape). Something similar may happen in the ramp paradigm. Here, the colors of both the cube and the ramp jointly determine where exactly

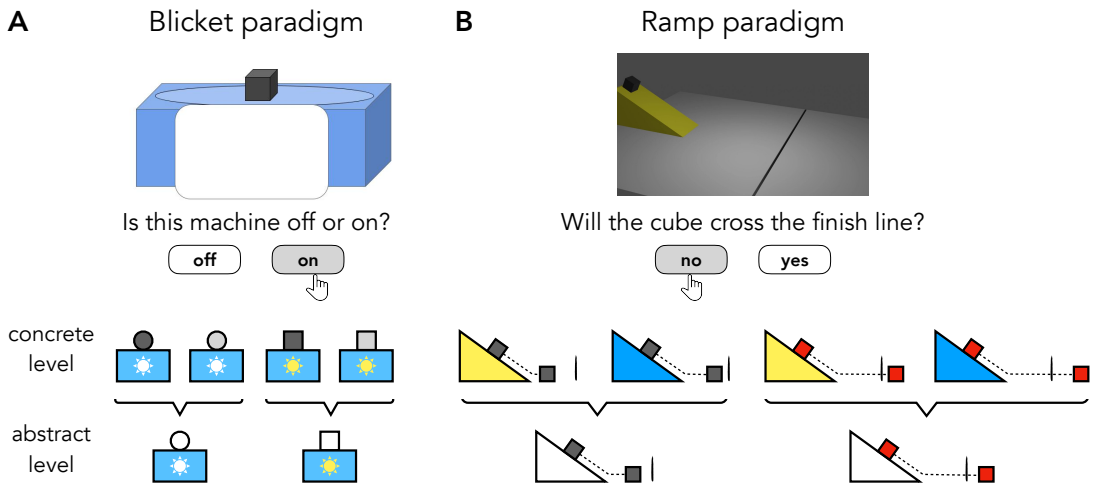

Figure 1

*Illustration of the blicket paradigm (**A**) and ramp paradigm (**B**). In the blicket paradigm, participants predict if the blicket detector (its state hidden by an occluder) will turn on. In the ramp paradigm, participants predict if a cube will cross the finish line. Both paradigms feature two sets of features that affect the outcome. In the examples shown, only one feature is key for correct prediction. In the blicket task, the object's shape determines if the detector activates, while its color is irrelevant. In the ramp paradigm, the cube's properties determine if it crosses the finish line; the ramp's properties do not. The causal abstraction model posits that people represent information at the level of abstraction best suited to their goal. Consequently, they may not remember details such as the blicket's color or the ramp's color if these are not relevant to the task.*

the cube ends up. But if all that matters is whether it crosses the finish line, then a learner might only pay attention to the cube and disregard the ramp.

We use these two paradigms to examine whether people learn abstract causal models. Our experiments feature a 'prediction task' where participants learn to predict what happens (just like shown in Figure 1). We then surprise them with a new task. In Experiment 1 (which uses the blicket paradigm), we ask them what the last trial looked like. We find that they are more likely to remember goal-relevant features (e.g., the shape of the cube rather than its color). In Experiments 2 and 3 (which both use the ramp paradigm), we surprise participants with a task where they have to predict exactly where the cube will end up (rather than merely predicting on which side of the finish line it ends up). We find that they make systematic errors: they know on which side of the line the cube will end up but not the exact position. Experiment 3 also asks participants to predict what would happen in novel situations. We find that what abstract causal model participants learned determines how they generalize to new situations. We develop a causal abstraction model and show that it accurately captures participants' judgments in the generalization task, and that their judgments cannot be explained by a feature-based model.

Experiment 1: Causal abstraction of blickets

In this experiment, we use the blicket detector paradigm (Figure 1a) to investigate whether people learn abstract representations that are tailored to their goal.

Methods

All experiments were pre-registered on the Open Science Framework, including information about the desired sample size, hypotheses, and statistical analyses. You can access all the pre-registrations, data, and materials here: https://github.com/cicl-stanford/causal_stories. All experiments were approved by Stanford's Institutional Review Board (#48665, "Understanding causal cognition").

Participants

A total of 482 participants were recruited through Prolific (*age*: M = 38, SD = 14; *gender*: 267 female, 192 male, 12 non-binary, 11 other or no response; *race*: 368 White, 45 Asian, 30 Black, 27 Multiracial, 2 Hispanic, 1 Native, 1 White African, and 8 no response; *ethnicity*: 428 Non-Hispanic, 39 Hispanic, 15 no response) took part in the four experimental conditions (*feedback*: $N = 124$, *no feedback*: $N = 118$, *short*: $N = 120$, *conjunctive*: $N = 120$). All participants were based in the US, fluent in English, and had approval ratings of at least 95% with 10 or more prior submissions. A target sample size of 120 participants was selected for each of the four conditions.[1] Participants were compensated with a base payment plus a performance bonus based upon their accuracy in the 'prediction task'. The average compensation exceeded $14 per hour in each condition.

Procedure

The stimuli in this experiment showed a 'blicket detector' with one of four objects placed on the machine: a dark cube, a light cube, a dark cylinder, or a light cylinder (see Figure 2). The blicket detector turns on (as indicated by the sun turning yellow) whenever blickets are placed on the machine. What gave away whether something was a 'blicket' was either its shape, color (or the combination), and this was counterbalanced between participants. In Figure 2, shape is the relevant feature, while color is irrelevant. Here, cubes (of any color) are blickets while cylinders (of any color) are not. For other participants, color was relevant while shape was irrelevant.

At the beginning of the experiment, participants were familiarized with the 'blicket detector' and informed that some but not all objects would make the detector turn on. A comprehension check ensured that they understood how it worked. The main body of the experiment consisted of a set of 'prediction trials'. In each trial, participants were presented with an image of a blicket detector, with one of the four objects placed on the detector like shown in Figure 1a. The status of the detector (whether it was 'on' or 'off') was hidden by a white occluder. Participants were asked to indicate whether the detector was 'on' or 'off'. After responding, the occluder was removed, the status of the detector revealed, and

[1] We simulated datasets assuming probabilities of 0.6, 0.25, 0.1, and 0.05 (from left to right) for the four different choice options shown in Figure 2. We ran a multinomial regression and computed whether the difference between options 2 and 3 was statistically significant at $\alpha = 0.05$ (two-sided). A sample size of 120 achieved a power of 0.8.

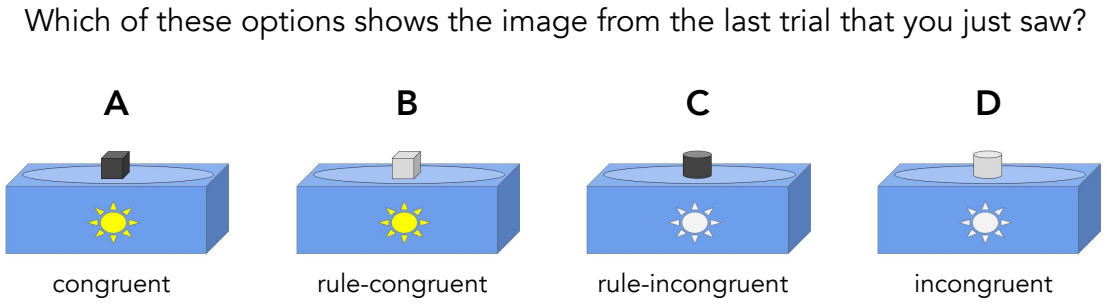

Figure 2

Experiment 1*. Example of the 'surprise task'. The cubes are blickets and the cylinders aren't. For the labels (which weren't shown to participants), we assume that the black cube was shown last in the 'prediction task' (see Figure 1a). In the 'no feedback' condition, the feedback was not revealed on the last test trial, so the options here looked like the one shown in Figure 1a (i.e., participants didn't see on the last trial whether the blicket was on or off).*

participants received feedback about whether their response was correct. Participants got a bonus for each correct response. After the last trial in the 'prediction task', participants viewed a 'surprise task' asking them to select which out of four images they had just seen (see Figure 2).

Design

The experiment had four conditions. In the 'feedback' condition, participants received feedback on the final prediction trial, indicating whether they responded correctly, before being presented with the surprise task. In the 'no feedback' condition, participants didn't receive feedback on the final trial. In the 'short' condition, participants only saw two prediction trials with the second one followed by the surprise task. The other conditions featured 16 prediction trials. Finally, in the 'conjunctive' condition, the blicket detector only turned on if two features were present (e.g., only black cubes were blickets). In each condition, we counterbalanced what features were causally relevant for the outcome.

Predictions

Figure 2 shows the different response options and corresponding labels in the 'surprise task'. We call responses 'congruent' when they correctly identify the object from the preceding trial (here, the black cube). Responses are 'rule-congruent' when they would have led to the same outcome as the correct response. Here, the light cube is rule-congruent because it is also a blicket. The black cylinder is 'rule-incongruent' – it shares its color with the correct response, but that feature is not the causally relevant one. We call the white cylinder 'incongruent' because it shares neither color nor shape with the correct response. Note that in the conjunctive condition, because only one object is a blicket, there are two

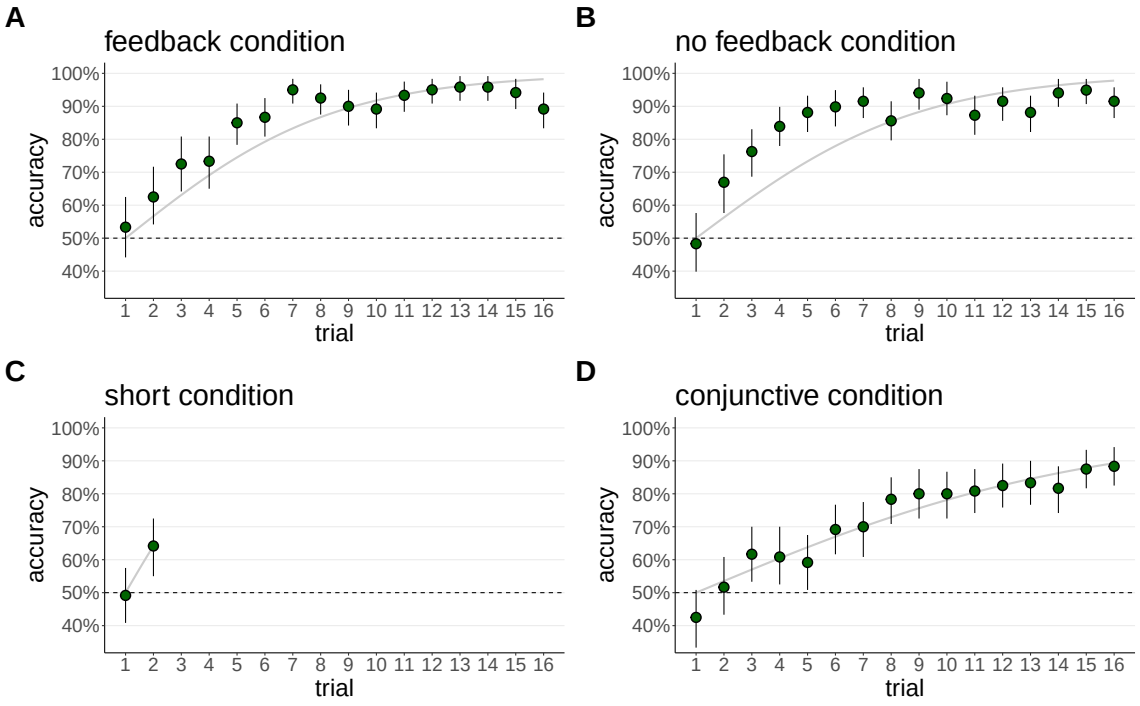

Figure 3

Experiment 1*. Accuracy in the prediction task separated by condition. Note: Error bars show 95% bootstrapped confidence intervals. Gray lines show logistic regression model fits with trial number as predictor. Dashed lines indicate chance level.*

partially matching responses (both the black cylinder and the light cube), and one incongruent option.

We predicted that, as participants learned what distinguishes blickets from non-blickets, they would begin to privilege the causally relevant information. For example, when the shape mattered and the color didn't, participants would be more likely to encode and remember the object's shape rather than its color. As a consequence, if participants made a mistake in recalling what they just saw, they would be more likely to select the rule-congruent rather than the rule-incongruent option (and least likely to select the incongruent option). In the 'feedback' condition, one might worry that participants would only remember the feedback they received (e.g., that the blicket detector was on) but not the object they had seen. This could explain why they would be more likely to choose the rule-congruent than the rule-incongruent option. To address this, we also included the 'no feedback' condition where participants didn't get feedback on the final prediction trial before the surprise task. If feedback was driving the effect, we would expect any potential difference in selections between the rule-congruent and rule-incongruent option to disappear in that condition. We predicted that in the 'short' condition, participants would not have enough evidence to learn what the causally relevant feature is, and would therefore be equally likely to select the rule-congruent and rule-incongruent options. In the 'conjunctive' condition, because both the color and the shape of the objects are causally relevant,

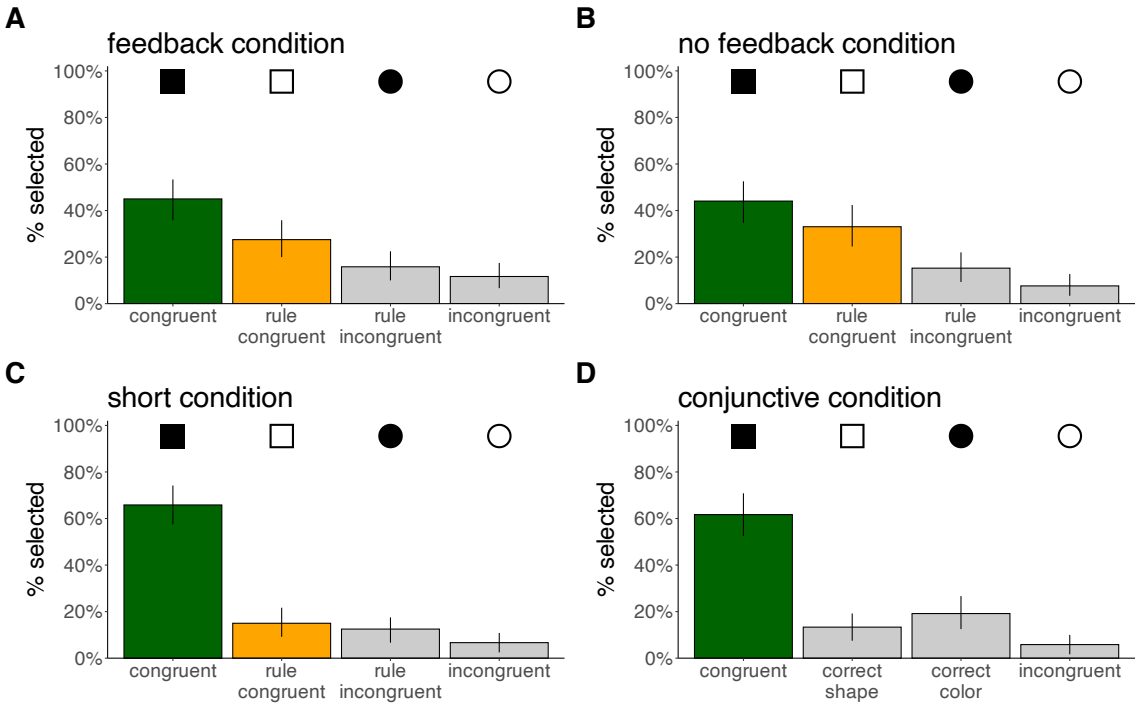

Figure 4

Experiment 1*. Participants' selections in the 'surprise task' separated by condition. Here, we assume that the object shown in the last prediction trial was a black cube, and that shape but not color is diagnostic of 'blickets' (see Figure 2). So, a rule-congruent response would be selecting the white cube, whereas a rule-incongruent response would be selecting the black cylinder. In the conjunctive condition, only black cubes were blickets. Note: Error bars show 95% bootstrapped confidence intervals.*

we predicted that participants would not have a privileged memory of either feature over the other, and thus would again be equally likely to select each partially congruent option.

Results

We discuss results from the 'prediction task' and 'surprise task' in turn.

Prediction task

Figure 3 shows participants' accuracy in the 'prediction task' across trials. Participants quickly learned which objects were blickets. In the conditions in which one feature mattered, more than 80% of participants successfully predicted whether the blicket detector was 'on' or 'off' by trial 5. In the conjunctive condition (Figure 3d), it took participants a little longer to learn the rule. In the short condition, participants' accuracy was just above 60% on the second (and final) trial.

Surprise task

Figure 4 shows participants' selections in the 'surprise task' for each condition. Results are aggregated over the counterbalanced conditions. Here, we assume that cubes are blickets and that the black cube was shown immediately prior to the surprise task. Our main hypothesis was that if people misremembered what they had just seen, they'd be more likely to select the rule-congruent option than the rule-incongruent option. In the 'feedback' condition, where participants saw the outcome on the final prediction trial, participants were more likely to select the rule-congruent than the rule-incongruent option but the difference was not significant, $\beta = 0.55$, 95% CI $[-0.01, 1.12]$, $p = .06$. In the 'no feedback condition', where participants didn't see the outcome on the final prediction trial, participants were significantly more likely to select the rule-congruent option $\beta = 0.77$, 95% CI $[0.21, 1.33]$, $p = .01$.

In the 'short condition' we correctly predicted that there would be no significant difference in selecting the rule-congruent versus rule-incongruent option, $\beta = 0.18$, 95% CI $[-0.5, 0.87]$, $p = .6$.[2] Participants were more likely to respond accurately here compared to the longer conditions. Finally, in the 'conjunctive' condition, we correctly predicted that there would be no difference in selecting either of the two partially matching responses, $\beta = -0.36$, 95% CI $[-1, 0.28]$, $p = .26$. Further, we correctly predicted that participants would be more likely to select an option which shared one feature with the correct response than the incongruent option, $\beta = 1.72$, 95% CI $[0.91, 2.52]$, $p = .0001$.

The results in Figure 4 aggregate over situations in which the last prediction trial featured a blicket or a non-blicket. As Table 1 shows participants were generally more accurate in recalling the features of blickets compared to non-blickets.

Discussion

Participants had no trouble learning what distinguished blickets from non-blickets. However, learning this simple causal rule had a consequence: participants learned to ignore some of the information. This was revealed through systematic errors they made when

Table 1

Experiment 1. *Percentage of congruent (= correct) responses depending on whether the 'surprise task' featured a blicket or not, separately for each condition. Participants were more likely to respond correctly when the last trial featured a blicket.*

	condition			
blicket	feedback	no feedback	short	conjunctive
yes	52%	47%	75%	77%
no	38%	42%	57%	57%

[2] Although not pre-registered, we also found that the proportion with which participants selected the rule-congruent option was significantly lower in the 'short' condition (15%) compared to the 'no feedback' condition (33%), Z ratio $= -3.20, p = .004$. This difference was marginally significant when comparing the 'short' with the 'feedback' condition (28%), Z ratio $= -2.34, p = .05$.

asked to recall what they had just seen. For example, when shape (but not color) was the causally relevant feature, participants' incorrect responses were more likely to have the same shape as the correct response, than they were to have the same color. The strength of this effect was striking. While 45% of participants selected the correct item when asked what they had just seen in the 'feedback' and 'no feedback' conditions, 30% of participants selected the incorrect but rule-congruent item.

Participants' tendency to recall the rule-congruent rather than the rule-incongruent option cannot be explained by them having remembered the feedback they received on the final prediction trial. We found an even stronger effect once we removed the feedback from the last prediction trial. One remaining possibility is that, while participants didn't receive feedback, they may still have remembered what response they produced on the last trial (whether they had clicked the 'yes' or 'no' button) and then chose an option that was consistent with their response. We address this concern in Experiment 2.

When participants weren't given enough evidence to learn what feature was causally relevant in the 'short' condition, they were more likely to correctly select the object they had last seen, and when they made a recall error, they were just as likely to select rule-congruent and rule-incongruent objects. Put differently, when participants had ample time to learn (in the feedback and no-feedback conditions), they performed markedly *worse* on the surprise test. In the 'conjunctive' condition, when both shape and color mattered, participants were more accurate in recalling the object they had last seen, and were equally likely to select either of the partially congruent options.

Participants were more likely to recall what happened when the object on the last trial was a blicket. This is not surprising in the 'conjunctive' condition (because only one of the four objects is a blicket), but more so in the other conditions (where two out of four objects are blickets). Notice also that this effect of better recalling a blicket was smaller in the 'no feedback' condition.

Experiment 2: Causal abstraction of physics

Experiment 1 provided evidence for the idea that people abstract away irrelevant information in a simple causal learning task. In Experiment 2, we extend these findings to the domain of intuitive physical reasoning (Wu et al., 2015). This time, we asked participants whether a cube sliding down a ramp will cross a finish line (see Figure 1b). Like in Experiment 1, we manipulated two features: the color of the cube and the color of the ramp. Each feature was diagnostic for the friction of that object. For example, in Figure 1b, the black cube has more friction than the red cube, and the yellow ramp has more friction than the blue one.

This physical setting expands the 'blicket detector' paradigm in several ways. First, on a fine level of granularity, the outcome now features four possible states: the different positions of where the cube ends up (see Figure 1b). On a more abstract level, however, there are only two outcome states that matter: whether or not the cube crosses the finish line. This allows us to test whether the use of causal models leads to compression, whereby people divide a large space of possible outcomes into a few discrete categories that matter for their goal. Second, this task addresses the concern from Experiment 1 that participants answered by remembering their own response immediately preceding the surprise task. Instead of asking participants what they saw on the last trial, this time we ask participants

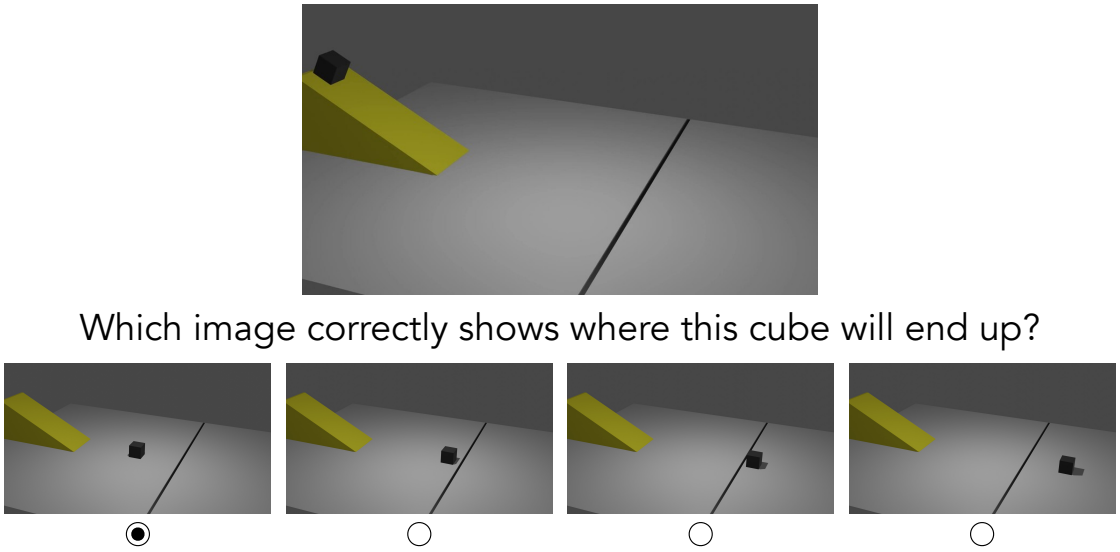

Figure 5

Experiment 2*. Example of the 'surprise task'. Participants' task was to select which image shows where the cube will end up. There were four tests like this for each combination of cube color and ramp color, with the order of trials randomized.*

to make predictions about where exactly the cube will end up (see Figure 5). This version of the 'surprise task' has the added benefit that we get more data from each participant: four judgments instead of one. Finally, by showing physically realistic animations of what happens in each scenario, this setting comes a little closer to the kinds of situations we may experience in our everyday lives.

Methods

Participants

Participants were recruited through Prolific using the same inclusion criteria as in Experiment 1. 359 participants (*age*: M = 38, SD = 15; *gender*: 186 female, 161 male, 9 non-binary, 3 no response; *race*: 272 White, 32 Black, 29 Asian, 18 Multiracial, 3 Native, 5 other or no response); *ethnicity*: 316 Non-Hispanic, 30 Hispanic, 13 no response) took part in three experimental conditions (*long*: $N = 120$, *short*: $N = 120$, *conjunctive*: $N = 119$). No one participated in more than one experiment or condition. The average compensation exceeded $11 per hour in each condition.

Procedure

The stimuli for this experiment consisted of videos showing a cube sliding down a ramp and then along a plane. Some cubes would slide beyond a finish line, while others would stop short. Cubes were either red or black and ramps were either blue or yellow. The color was diagnostic for its surface friction. After sliding down the ramp, a cube would stop in one of four equally spaced positions. The first and second positions were before the

finish line, and the other two positions after. The frictions of the objects were set such that either the cube friction or the ramp friction determined whether a cube would cross the finish line. For example, in Figure 1b, red cubes cross the finish line and black cubes don't. Here, cubes slide further on blue compared to yellow ramps. So, although both the ramp and the cube contribute to the cube's final position, attending to the cube color alone is sufficient for predicting whether it will cross the finish line.

Participants were instructed that they would need to make predictions about whether the cube would cross the finish line in each of the four possible scenarios. Participants were then given a brief comprehension check and informed that they would receive a bonus for each correct prediction. The main body of the experiment then consisted of a set of the 'prediction task' in which participants were presented with an image showing a cube at the top of a ramp like in Figure 1b, and asked if it would cross the finish line. After responding with 'yes' or 'no', participants were shown a video of the cube sliding down the ramp and coming to rest. They received feedback about whether their response was correct.

After completing the final trial of the 'prediction task', participants saw a 'surprise task'. Participants were asked: "Which image correctly shows where this cube will end up?" like in Figure 5. They responded by selecting one of the four images. The 'surprise task' included one trial for each of the four scenarios (black/red cube on a yellow/blue ramp). The order of these trials was randomized, and participants received no feedback.

Design

The experiment had three conditions. In the 'long' condition, participants completed 16 trials in the prediction task. In the 'short' condition, participants completed only 4 trials (one for each combination of cube and ramp). Finally, in the 'conjunctive' condition, the finish line was moved such that it fell between the third and fourth position. As a result, whether the cube crossed the finish-line now depended on both the cube and ramp friction. In the 'conjunctive' condition participants completed 16 trials. We counterbalanced whether the cube type or the ramp type were causally relevant, as well as what colors were diagnostic for low and high friction.

Predictions

We pre-registered the following predictions about participants' selections in the 'surprise task'.

In the 'long' condition, we predicted that participants' selections across all four trials would be more strongly affected by the feature that was causally relevant for their task (e.g., the color of the cube) compared to the irrelevant feature (e.g., the color of the ramp). We also predicted that for the subset of trials in which the correct response was position 2 or position 3, participants would be more likely to choose the outcome-congruent response than the outcome-incongruent response (see Figure 5). For example, if the correct response was 2, participants would be more likely to select 1 than 3. While both of these positions are equidistant from position 2, they fall on different sides of the finish line. These results would be consistent with the idea that people use causal models to reduce the dimensionality of the outcome space.

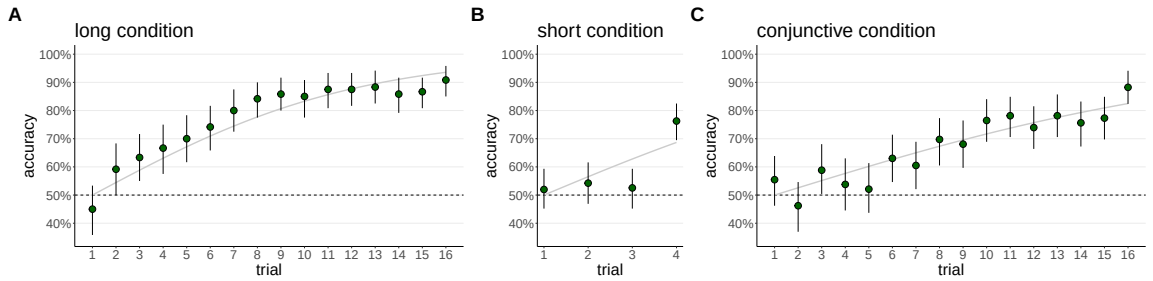

Figure 6

Experiment 2. *Accuracy in the prediction task separated by condition.* Note: *Error bars are 95% bootstrapped confidence intervals. Gray lines show logistic regression model fits with trial number as predictor. Dashed lines indicate chance level.*

In the 'short' condition, we predicted that there would be no difference in how strongly the causally relevant and irrelevant feature affected participants' selections, and that they would be just as likely to select outcome-congruent or outcome-incongruent responses when the cube's correct final position was 2 or 3. These results would suggest that one needs enough data to learn the relevant causal model first in order to compress the outcome space.

Finally, in the 'conjunctive' condition, we predicted that, when the correct position was 3, participants would be more likely to selection position 2 (which would lead to the same outcome) than position 4. We also predicted that participants would be more likely to select the correct response when the cube's final position was 4 compared to any of the other positions. These results would be consistent with the idea that people learn a causal model that's suited to their goal and that the outcome space is compressed accordingly.

Results

We discuss the results of the 'prediction task' and the 'surprise task' in turn.

Prediction task

Figure 6 shows participants' accuracy in predicting whether the cube would cross the finish line over the course of the 'prediction task' trials. Somewhat surprisingly, participants' accuracy in the short condition was quite high on the fourth and final trial. Like in Experiment 1, participants found it more difficult to learn the conjunctive rule, but achieved similar accuracy by the end.

Surprise task

We discuss participants' responses separately for each condition.

Long condition. Figure 7a (top) shows participants' selections in the 'long' condition for each of the four combinations of cubes and ramps. Here, we assume that the red cube crosses the finish line whereas the black cube doesn't, and that cubes slide further on blue than yellow ramps. The green bars in Figure 7 show correct responses, and the orange

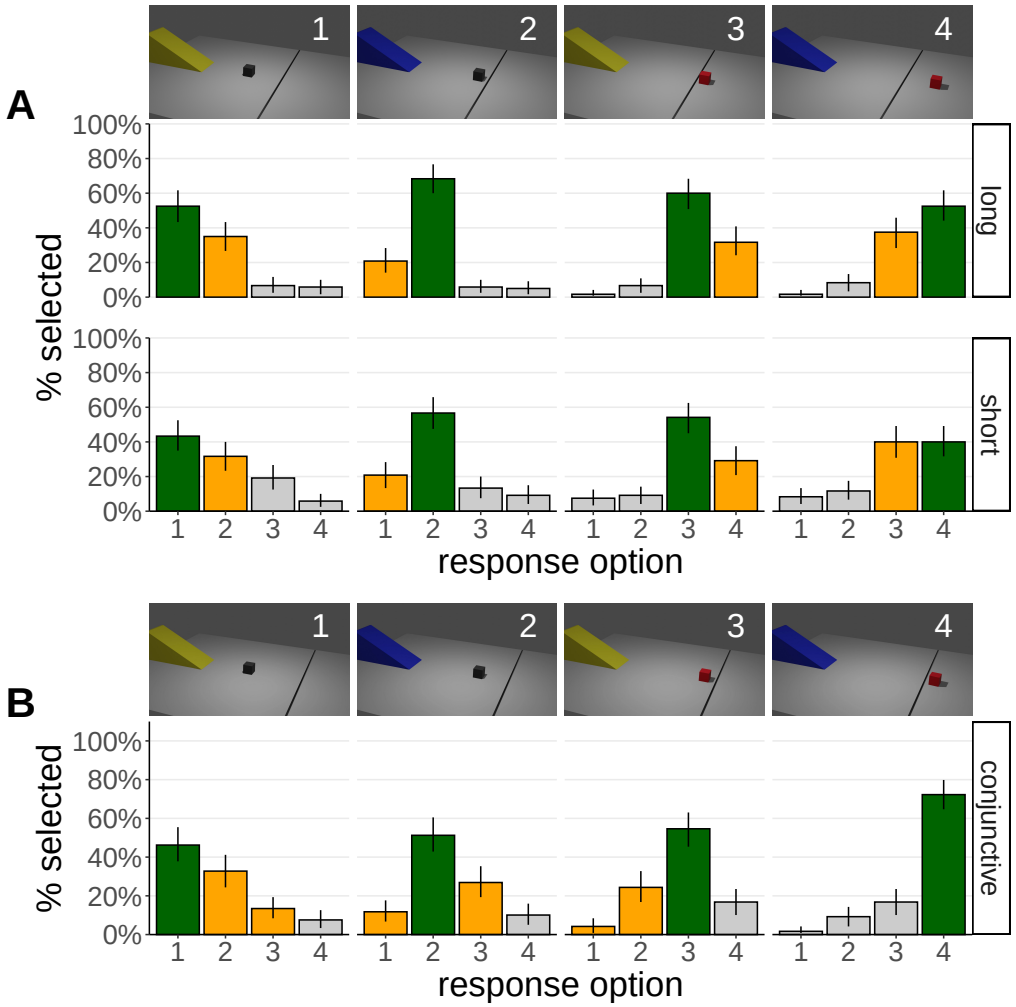

Figure 7

Experiment 2*. Participant selections of different end positions in the four trials of the 'surprise task'. Green bars indicate the correct response, orange bars show outcome-congruent responses.* Note*: Error bars are 95% bootstrapped confidence intervals. Notice that we counterbalanced which feature mattered for whether a cube ended up on the right side of the finish line (ramp vs. cube). We also counterbalanced the mapping of the colors to the features (e.g., black vs. red being more friction). In this figure, we show combined data from all counterbalanced conditions, and illustrate the results based on the scenario where the cube feature matters, where black cubes have more friction than red cubes, and where yellow ramps have more friction than blue ramps.*

bars show incorrect but outcome-congruent responses. For example, when the correct response is that the cube would end up in position 2, the outcome-congruent response would be position 1 because for both positions, the cube would not have crossed the finish line.

To test the prediction that the causally relevant feature affected participants' selec-

tions more strongly than the irrelevant feature, we ran a Bayesian ordinal mixed effects regression with 'relevant' and 'irrelevant' feature plus their interaction as fixed effects, and random intercepts for participants.[3] We then computed a distribution of the difference between the posterior on the 'relevant' and the 'irrelevant' predictor to test whether the relevant feature mattered more. As predicted, participants' selections were more strongly affected by the 'relevant' than the 'irrelevant' feature, $\beta = 0.85$, 95% Credible Interval (CrI) = $[0.7, 1.01]$.

To test the prediction that participants would be more likely to select outcome-congruent responses when the correct position was 2 or 3, we ran a Bayesian mixed effects logistic regression with an intercept as fixed effect as well as random intercepts for participants. We coded outcome-congruent responses as '1' and outcome-incongruent responses as '0'. As predicted, participants were more likely to select outcome-congruent options than outcome-incongruent ones, 90% [75%, 99%].

Short condition. Figure 7a (bottom) shows participants' selections in the 'short' condition. In contrast to what we predicted, the causally 'relevant' feature again had a stronger influence on participants' selections than the 'irrelevant' feature, $\beta = 0.53, 95\%\ CrI = [0.39, 0.67]$, though this difference was smaller than in the 'long' condition. Similarly, against our prediction, participants were more likely to select the outcome-congruent response when the final cube position was 2 or 3, 73% [60%, 88%], but this effect was also weaker than in the 'long' condition.

Conjunctive condition. Figure 7b shows participants' selections in the 'conjunctive' condition. Notice that the images are different here because now only cubes that reach position 4 cross the finish line. When the ground truth position was 3, participants were more likely to select the outcome-congruent position 2 than position 4 (59% [46%, 72%]) but, against what we predicted, the credible interval of the estimate did not exclude 50%. As predicted, participants' accuracy for position 4 (81% [71%, 90%]) was greater than that for the other three positions (51% [41%, 62%], $\beta = 1.44$ [0.86, 2.01]).

Discussion

Experiment 2 again shows a pattern of results that is consistent with the idea that participants learn abstract models that privilege causally relevant information. As a consequence, participants produced systematic errors when confronted with a 'surprise task' for which their abstract causal model was inadequate. Participants were more likely to recall incorrect outcomes that were consistent with the abstract causal rule that they had learned. Participants produced these errors even though they had ample experience. In the 'long' and 'conjunctive' conditions, participants viewed each of the four clips four times in the prediction task.

Unlike what we predicted, and unlike what we found in Experiment 1, participants made systematic errors even when they had relatively little experience in the 'short' condition. It's possible that participants were able to build the relevant causal abstraction fairly

[3] We pre-registered Bayesian analyses for Experiment 2 because it was easier to implement ordinal mixed effects regression models this way. All Bayesian models were written in `Stan` (Carpenter et al., 2017) and accessed with the `brms` package (Bürkner, 2017) in `R` (R Core Team, 2019). We consider a prediction confirmed when the 95% credible interval of the posterior estimate for the predictor of interest excludes 0 (or 50% depending on the statistical model).

quickly in this task. On the fourth trial, participants already had an accuracy of almost 80%. Notice that unlike in Experiment 1 (where participants only viewed two trials that were randomly selected), this time participants saw trials with each of the four possible combinations of the two features.

In the conditions in which one feature was causally relevant, participants were more accurate when the final position was close to the finish line (long condition: 64%; short condition: 55%) than when it was further away (long condition: 52%; short condition: 42%). In all three conditions, participants were very unlikely to select a response that was inconsistent with the outcome, such as selecting position 3 when the correct position was 2 (in the 'long' or 'short' conditions). This suggests that participants may have paid particular attention when the outcome was close.

In the 'conjunctive' condition, participants viewed essentially the same video clips (with the finish line moved one position forward) but produced very different responses. For example, when the end position was 2, participants were now more likely to think that it was position 3 than position 1. This is the opposite error pattern from the other two conditions. This illustrates how people learned causal abstractions that were specifically suited to their goals.

Experiment 3: Constructing causal stories of what happened

Experiment 2 showed that participants make systematic errors in a physical prediction task. They were able to say on which side of a finish line a cube would end up, but not where exactly. This result is in line with the idea that people construct abstract causal models that prioritize goal-relevant information. However, it's also possible that participants learned a simple (and non-causal) mapping from features to outcomes. For example, they may have simply learned that black cubes cross the line and that red cubes don't without inferring anything about the underlying physical parameters, such as the friction of the cubes and ramps. We made several changes in Experiment 3 that allow us to better differentiate what participants would do if they learned goal-dependent causal abstractions of the task as compared to direct mappings from observable features to outcomes.

First, we no longer show animations of how the cubes slide down the ramp. Participants only see an image of the initial state and then, after predicting whether the cube will cross the finish line, an image of the final state. Removing the animation creates ambiguity about what happened. As we will see below, the causal abstraction model is sensitive to the causal process that connects initial and final states, whereas the feature-based model is not.

Second, we added a condition where the ramp faces the other direction (see Figure 8). In both the 'ramp forward' and the 'ramp backward' condition, the cube always ends up on the right side of the ramp in the 'prediction task'. Intuitively, in the 'ramp forward' condition, a simple explanation of how the cube ended up where it was is that it slid down the ramp. That explanation doesn't work in the 'ramp backward' condition. To explain how the cube ended up on the right side of a backward facing ramp, one might consider that someone pushed the cube in that direction, or the ramp has a mechanism that projects the cube in that direction. The causal abstraction model infers what happened based on the data, but the feature-based model doesn't – it just learns a mapping between features of the initial state and the final state.

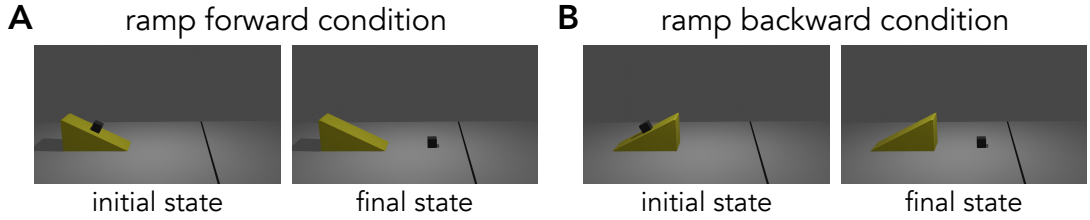

Figure 8

Experiment 3*. Example stimuli from the two experimental conditions: 'ramp forward' (**A**) and 'ramp backward' (**B**). This time, participants didn't see animations of the cube moving. Instead, they only viewed the initial state and, after responding, saw the final state. Note that in both the 'ramp forward' and the 'ramp backward' conditions, the cube ends up on the right side of the ramp. In the 'ramp forward' condition, a plausible explanation of what happened is that the cube slid down the ramp. In the 'ramp backward' condition, one possible explanation is that the cube was pushed up the ramp somehow.*

Third, after the 'surprise task' (which is the same as in Experiment 2), participants do a 'generalization task'. The example in Figure 9 features a gray ramp that participants haven't seen before. Their task is to predict where the two cubes (which they have seen before) will end up. For participants in the 'ramp forward' condition, the prediction should be straightforward: the cubes are going to end up on the right side of the ramp (consistent with the explanation that cubes slide down ramps). But what about participants in the 'ramp backward' condition? Are they going to predict that the cubes will end up on the left side of the ramp (which would be consistent with a mechanism inside the ramp that projects the cubes forward), or on the right side (which would be consistent with the idea that someone pushes the cubes toward the finish line)? If participants came up with a story of how the data was generated, then that story might influence how they generalize to new situations. In the experiment, we show four different generalization trials that manipulate the ramp direction, and whether the cubes or the ramp are unknown (see Figure 10). The causal abstraction model and the feature-based model make different predictions on this task, and we will describe each model in more detail below.

Methods

Participants

239 participants were recruited through Prolific (*age*: $M = 37$, SD $= 11$; *gender*: 137 female, 98 male, 2 non-binary, 2 other or no response; *race*: 158 White, 45 Black, 22 Asian, 7 Multiracial, 2 Native, 1 Hispanic, and 8 no response; *ethnicity*: 215 Non-Hispanic, 19 Hispanic, 5 no response) took part in the four experimental conditions (*ramp forward, cube matters*: $N = 59$, *ramp forward, ramp matters*: $N = 61$, *ramp backward, cube matters*: $N = 62$, *ramp backward, ramp matters*: $N = 57$).

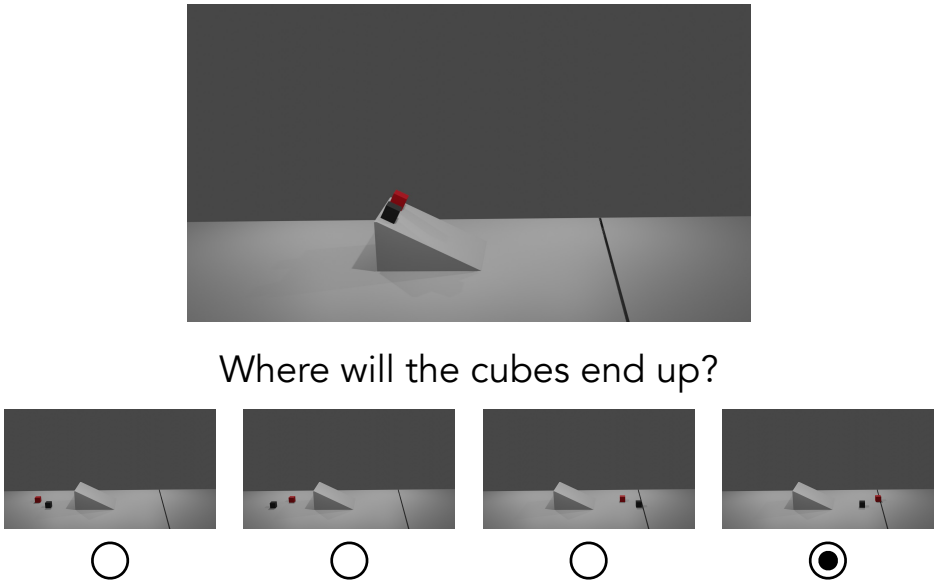

Figure 9

Experiment 3*. Example trial in the 'generalization task'. Participants are asked to select the image that shows where the cubes will end up. In this case, the participant selected the image on the right.*

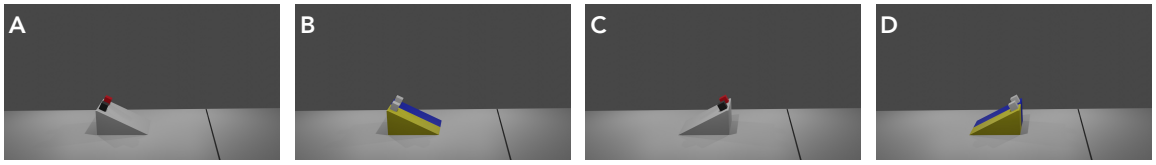

Figure 10

Experiment 3*. The four test trials in the* generalization task*. Participants see each of these four configurations, and they have to select where they think the cubes will end up. Figure 9 shows the four response options for trial A which features two different cubes and the ramp facing forward. The response options were similar for the other three trials. In two of the images the cubes were on the left side of the ramp, and in the other two, on the right side. For each pair of images on either side, one cube was further away from the ramp in one image, and the other cube in the other image.*

Design

We manipulated two factors in the experiment: (1) the ramp orientation in the 'prediction task', and (2) what the causally relevant feature was. The ramp was either oriented forward (see Figure 8a) or backward (see Figure 8b). The relevant feature was either the cube color or ramp color. For each of the four conditions, we counterbalanced the mapping between color and outcome. For example, for half of the participants it was the red cube that crossed the line, and for the other half it was the black cube (similarly

for the yellow and blue ramp).

Procedure

The procedure was largely identical to that of Experiment 2. However, one important difference was that in the 'prediction task', participants no longer saw video animations. Instead, they only saw the initial image, made their prediction, and then saw the final image. Participants viewed 16 trials (4 of each type of trial) in randomized order. The 'surprise task' was the same as in Experiment 2. Participants saw the 4 surprise trials in randomized order. This time, we added a 'generalization task' as described above. Participants saw the four generalization trials shown in Figure 10. In the first two trials, the ramp orientation was the same as it had been in the 'prediction task' and the 'surprise task'. In the final two trials, the ramp orientation was reversed. For example, for a participant who was in the 'ramp forward' condition (see Figure 8a), the first two generalization trials would be the ones shown in Figure 10a and b, and the remaining trials would be those shown in Figure 10c and d. A participant in the 'ramp backward' condition, would see trials c and d before a and b. The order of presentation within each pair of trials was randomized. We counterbalanced whether the cube color or the ramp color was causally relevant for whether a cube ended up on the left or right of the finish line. We also counterbalanced which cube or ramp were shown in front for the generalization trials. For example, for some participants, the black cube was in front (like in Figure 9), and for others, the red cube.

Predictions

We will discuss the predictions for the 'surprise task' and 'generalization task' in turn.

Surprise task

Our predictions for the surprise task are the same as in Experiment 2's 'long condition'. We predicted that participants' selections would be more strongly affected by the causally relevant feature (e.g., the cube color) compared to the irrelevant feature (e.g., the ramp color). We also predicted that for the subset of trials in which the correct response was position 2 or position 3, participants would be more likely to choose the outcome-congruent response than the outcome-incongruent response. For example, if the correct response was 2, participants would be more likely to select 1 than 3.

Generalization task

Based on the idea that people learn abstract causal models that capture how the data was generated, we predicted that participants' responses would differ between the 'ramp forward' and 'ramp backward' condition. For generalization trials where the orientation of the ramp was different from what participants had seen before, participants in the 'ramp forward' condition would be more likely than participants in the 'ramp backward' condition to select response options where the cubes would end up on the opposite side. We also developed two models that yield quantitative predictions about participants' selections in the generalization task: a causal abstraction model, and a feature-based model. We will describe each model in turn.

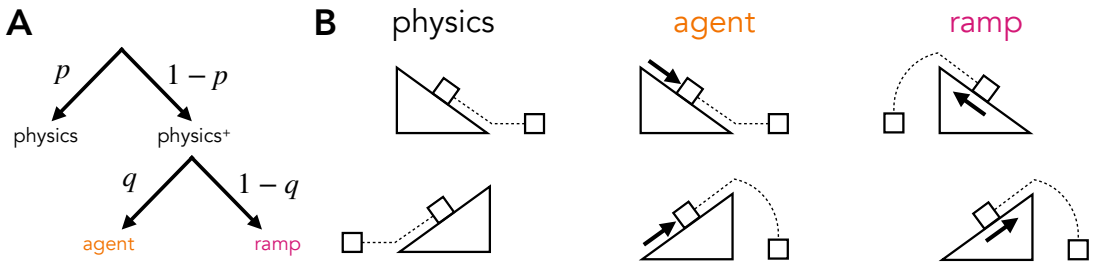

Figure 11

Model predictions for the 'generalization task'. ***A*** *Decision tree of what causal story to tell about how the data came about. We assume that people initially place a high probability p on the system being purely physical. However, upon receiving evidence that it's not (i.e., a cube ending up on the right side of the backwards facing ramp), people consider alternative explanations ('physics$^+$'), such as an 'agent' that pushes the cube toward the goal (with probability q), or that the 'ramp' pushes up the cube (with probability* $1 - q$*).* ***B*** *Model predictions of what will happen in the generalization task. In the 'physics' model, the cube just slides down the ramp. In the 'agent model', some external force pushes the cube toward the finish line (which is always on the right side of the ramp). In the 'ramp model', there is some internal force in the ramp that pushes the cube up the ramp.*

The causal abstraction model. We construe participants' task of figuring out what happened and generalizing to new situations as a Bayesian inference problem. Accordingly, the goal is to update one's prior beliefs about different hypotheses $p(\text{hypothesis})$ to a posterior belief $p(\text{hypothesis}|\text{data})$ by considering the likelihood of the observed data under each hypothesis $p(\text{data}|\text{hypothesis})$:

$$p(\text{hypothesis}|\text{data}) \propto p(\text{data}|\text{hypothesis}) \cdot p(\text{hypothesis}) \tag{1}$$

The causal abstraction model considers three different hypotheses of how the data may have been generated (see Figure 11). First, according to the 'physics' hypothesis, the cubes just slide down the ramp. Second, according to the 'agent' hypothesis, an (invisible) agent pushes the cube toward the finish line. Finally, according to the 'ramp' hypothesis, ramps push the cubes upward. Notice that the data that participants see in the 'ramp forward' condition is equally consistent with the 'physics' and the 'agent' hypothesis, but not with the 'ramp' hypothesis. Conversely, the data in the 'ramp backward' condition is equally consistent with the 'agent' and the 'ramp' hypothesis, but not with the 'physics' hypothesis.

To compute the posterior belief in a hypothesis, such as the 'physics' hypothesis, given observations of where the cube ended up, the model computes

$$p(\text{hypothesis}_{\text{physics}}|\text{cube position}) = \frac{p(\text{cube position}|\text{hypothesis}_{\text{physics}}) \cdot p(\text{hypothesis}_{\text{physics}})}{\sum_{i \in \{\text{physics,agent,ramp}\}} p(\text{cube position}|\text{hypothesis}_i) \cdot p(\text{hypothesis}_i)}, \tag{2}$$

where $p(\text{hypothesis}_{\text{physics}})$ is the prior probability assigned to the 'physics' hypothesis, and $p(\text{cube position}|\text{hypothesis}_{\text{physics}})$ is the likelihood of the observed cube position under that hypothesis.

Two parameters characterize the model's prior distribution over the three hypotheses: p and q (see Figure 11a). We assume that participants update their beliefs about the different hypotheses in light of the evidence they see in the prediction task. For simplicity, we assume a likelihood function $p(\text{cube position}|\text{hypothesis}_i)$ that assigns a likelihood of 1 if the cube position is consistent and ϵ otherwise. We take ϵ to be a small value that makes it such that a hypothesis is not fully ruled out even if the evidence appears to contradict it. For example, this will make it such that the 'physics' hypothesis would not fully ruled out even if a cube ended up the right of a backward facing ramp. As illustrated in Figure 11b, if the ramp faces forward, the 'physics' and 'agent' hypothesis assign a likelihood of 1 for cubes on the right (and ϵ on the left). The 'ramp' hypothesis does the opposite. If the ramp faces backward, the 'agent' and 'ramp' hypothesis assign a likelihood of 1 for cubes on the right (and ϵ on the left). Here, the 'physics' hypothesis does the opposite.

In the 'generalization task', we ask participants to select where the cubes would end up in a novel situation that they hadn't seen before (see Figure 9). To predict participants' selections in this task, the model computes the posterior predictive probability of the four response options

$$p(\text{option}_i|\text{data}) = \sum_{j \in \{\text{physics,agent,ramp}\}} p(\text{option}_i|\text{hypothesis}_j) \cdot p(\text{hypothesis}_j|\text{cube positions}), \quad (3)$$

where $p(\text{option}_i|\text{hypothesis}_j)$ is the likelihood of the new data shown in response option_i under hypothesis_j (see, e.g., Figure 9) and $p(\text{hypothesis}_j|\text{cube positions})$ is the posterior probability of that hypothesis based on the data seen so far (computed in Equation 2).

We compute $p(\text{option}_i|\text{hypothesis}_j)$ via simulation using the physics engine with which we generated the stimuli. We assume that participants are uncertain about two properties when simulating where a cube would end up under a given hypothesis: the friction of the ramps, and the friction of the cubes. To capture this uncertainty, the model adds Gaussian noise to the ground truth friction parameters before it simulates what would happen. σ_{cube} and σ_{ramp} determine how much noise is added to the cube and ramp friction, respectively. To calculate the likelihood of an observed cube position under each hypothesis, the model runs m noisy simulations. Each simulation yields a final position for each of the two cubes. Based on these simulations, we compute a probability distribution over the final cube positions using a Gaussian kernel. This method has one free parameter for setting the bandwidth of the Gaussian kernel, which we denote by κ. To compute $p(\text{option}_i|\text{hypothesis}_j)$ for any given hypothesis, we then multiply the probability of each of the two cubes ending up where they did, according to that hypothesis.

Finally, to translate the posterior predictive distribution in Equation 3 into a discrete choice, the model uses a softmax function

$$p(\text{option}_i) = \frac{\exp^{\beta \cdot p(\text{option}_i|\text{data})}}{\sum_{j=1}^{4} \exp^{\beta \cdot p(\text{option}_j|\text{data})}}, \quad (4)$$

where β determines the extent to which the model chooses the option with the highest posterior predictive probability. For large values of β, the model deterministically chooses the best option (i.e., it hard-maxes). If β was 0, the model would choose randomly.

Overall, this model has seven free parameters. p and q determine the prior distribution over hypotheses, ϵ determines the likelihood of observing events that are inconsistent

with a given hypothesis, σ_{cube} and σ_{ramp} determine how much noise is added to the physical simulations, κ determines the width of the Gaussian kernel that turns the simulation outcomes into a probability distribution, and β determines how the continuous posterior predictive distributions are turned into a discrete choice. We fitted these parameters by maximizing the likelihood of participants' generalization choices. The best-fitting values are: $p = 0.91$, $q = 0.34$, $\epsilon = 0.05$, $\sigma_{\text{cube}} = 0.1$, $\sigma_{\text{ramp}} = 1.6$, $\kappa = 0.6$, and $\beta = 341910$.[4]

A feature-based model. As an alternative model for how people might generalize what they've learned to new situations, we consider a feature-based model. This model does not infer a causal story about how the data was generated, but instead uses a set of four features (and their interactions) to predict participants' selections. The `orientation` feature encodes whether the ramp faces forward or backward in the 'prediction task'. The `side` feature encodes whether the cubes are on the side of the ramp that's consistent with that in the 'prediction task'. For example, in the 'ramp forward' condition, this feature labels the cubes as 'consistent' when they are shown on the right side when the ramp faces forward (e.g., options 3 and 4 in Figure 9) and 'inconsistent' when they are on the left side (options 1 and 2). Conversely, when the ramp faces backward, the 'consistent' options would be the cubes on the left side, and the 'inconsistent' ones on the right side. In the 'ramp backward' condition, all of this flips – now, cubes on the right side of a backward facing ramp are 'consistent' and so on. The `property` feature encodes whether the property that participants are asked to generalize was causally 'relevant' or 'irrelevant' for whether the cube would end up on the right of the finish line in the 'prediction task'. In our running example, we assumed that the cube color was the relevant feature, and that the ramp color was irrelevant. Finally, the `friction` feature encodes whether the cube (or ramp) friction is consistent with what participants have learned in the 'prediction task'. For example, let's assume that in Figure 9 the black cube has higher friction than the red one. Then, the 'consistent' options would be the ones that show the black cube closer to the ramp and the red one further away (options 1 and 4). The 'inconsistent' options show the black cube further away from the ramp and the red one closer (options 2 and 3).

To fit this model to the data, we computed a linear regression of the following form:

$$p(\text{option}_i) = \alpha + \beta_1 + \text{orientation} + \beta_2 \cdot \text{side} * \beta_3 \cdot \text{property} * \beta_4 \cdot \text{friction}, \tag{5}$$

where α is the intercept, β_1 is the coefficient for the orientation feature, β_2 for the `side` feature, β_3 for the `property` feature, and β_4 for the `friction` feature. The '*' indicates the interaction between the features (including two-way and three-way interactions). This model has 16 free parameters in total.

Results

We discuss the results of the 'prediction task', 'surprise task', and 'generalization task' in turn.

Prediction task

Figure 12 shows participants' accuracy in the prediction task, separately for the 'ramp forward' condition (Figure 12a) and the 'ramp backward' condition (Figure 12b).

[4] The beta parameter is large because the softmax in Equation 4 uses values from the posterior predictive distribution as inputs (see Equation 3) and these values are very small.

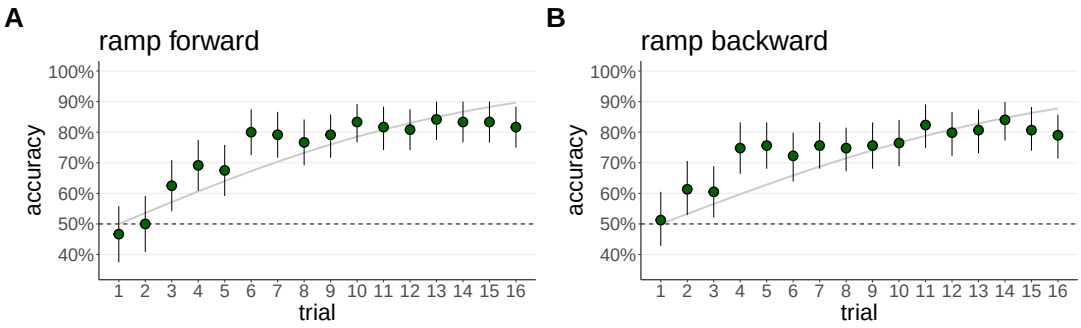

Figure 12

Experiment 3*. Accuracy in the prediction task in the 'ramp forward' (**A**) and the 'ramp backward' condition (**B**) See Figure 8 for a visualization of the setup.* Note*: Error bars show 95% bootstrapped confidence intervals. Lines show logistic regression model fits with trial number as predictor.*

There was no credible difference in accuracy between the two conditions, 0.02 [−0.28, 0.33]. By the end of the prediction trials, participants' accuracies in both conditions were around 80%. This is similar to what we saw in the 'long condition' in Experiment 2. This suggests that viewing animations of the cube sliding down the ramp did not affect participants' accuracy in this task. Further, it's no more difficult for participants to accurately learn whether cubes end up on the right side of the finish line when the ramp faces backward compared to when it faces forward.

Surprise task

Figure 13 shows the results of the 'surprise task' separately for whether the ramp faced forward (Figure 13a) or backward (Figure 13b). As predicted, across both conditions, participants' selections were more strongly affected by the causally relevant feature compared to the irrelevant feature, $\beta = 0.7$, 95% CrI $= [0.6, 0.8]$. We again found that participants were more likely to select incorrect responses that were outcome-congruent than ones that were outcome-incongruent when the final position was 2 or 3, 96% [85%, 99%].

Overall, the results in both the 'ramp forward' condition and the 'ramp backward' condition were remarkably similar to each other, and to the results in Experiment 2, with one notable difference. In Experiment 2, we found a stronger differentiation between the true response (show in green in Figure 7) and the outcome-congruent incorrect response (shown in orange), when the correct position was 2 or 3, compared to when it was 1 or 4. This effect was not replicated here. Instead, we found that participants were equally likely to confuse position 1 and 2, or 3 and 4, no matter what the ground truth position was. It's possible that having seen the video of the cube sliding in Experiment 2 highlighted the situations in which the cube ended up closer to the finish line (positions 2 or 3), compared to when it was further away (positions 1 or 4).

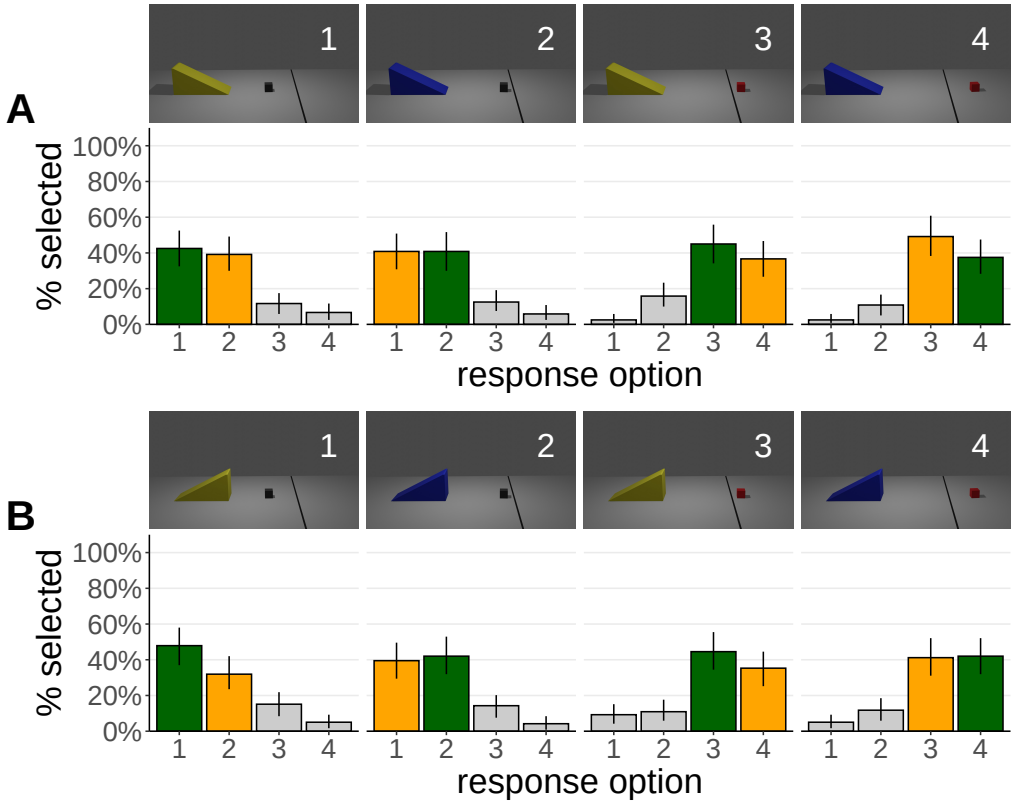

Figure 13

Experiment 3*. Participant selections of different end positions in the four trials of the 'surprise task' in the 'ramp forward' condition (**A**), and 'ramp backward' condition (**B**). Green bars indicate the correct response, orange bars show outcome-congruent responses.* Note*: Error bars are 95% bootstrapped confidence intervals. Notice that just like in Figure 7, we show the results aggregated across all counterbalanced conditions.*

Generalization task

Figure 14 shows the selections from the 'ramp forward' condition (Figure 8a). In Figure 14a, most participants thought that the cube would end up on the right side of the ramp, and that the red cube would end up further from the ramp than the black cube. In Figure 14b, participants again thought that the cubes would end up on the right side, but this time, a roughly equal number of participants thought that the gray cube on the blue ramp would go further versus the gray cube on the yellow ramp. This result reflects two things: first, almost all participants believed that the cubes would end up on the right side of the ramp, just like they had seen throughout the experiment so far. Second, these results are consistent with the idea that participants had learned more about the causally relevant feature (here, the cube friction) than the irrelevant feature (here, the ramp friction).

In Figure 14c, where the ramp was now flipped, most participants thought that the cubes would end up on the left side of the ramp, and that, again, the red cube would go

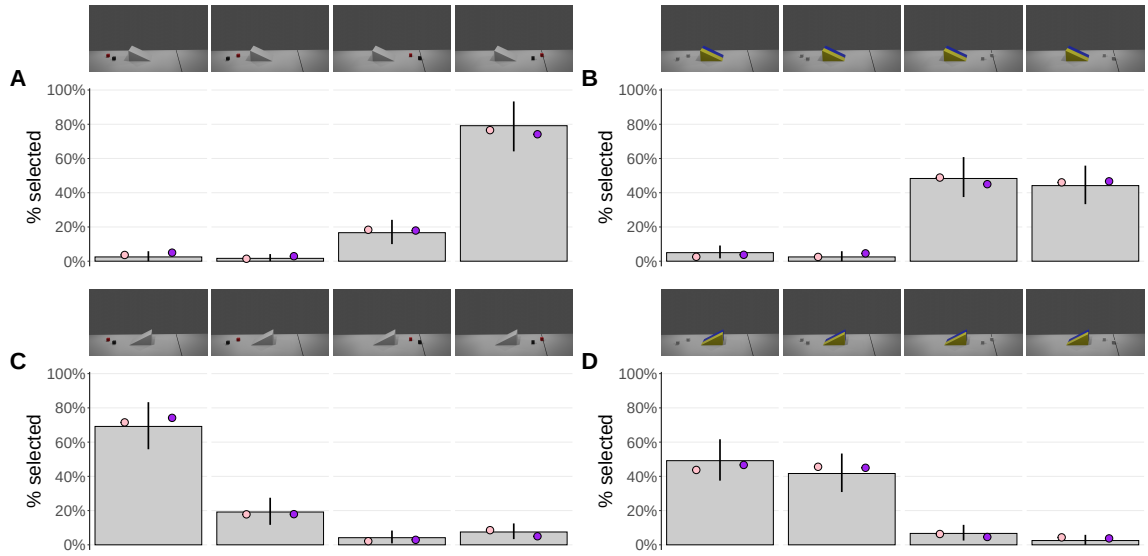

Figure 14

Experiment 3. *Participant selections in the 'generalization task' in the 'ramp forward' condition together with the predictions of the causal abstraction model (○) and the feature-based model (●).* Note*: Error bars are 95% bootstrapped confidence intervals. Notice that the results here are shown based on the combined data across all counterbalanced conditions, and illustrated for the case in which the cube featured mattered (rather than the ramp feature). Here, the black cube has more friction than the red cube, and the yellow ramp has more friction than the blue ramp.*

further than the black cube. In Figure 14d, participants thought that the cubes would end up on the left side but didn't know which cube would go further. Overall, the results in the generalization task for the backward ramp closely mirror those for the forward ramp. Notice that, so far, both the causal abstraction model and the feature-based model can account for this set of results (albeit assuming different underlying cognitive mechanisms).

The causal abstraction model places a greater prior belief in the 'physics' explanation (see Figure 11b) – that cubes simply slide down ramps. Based on the data that it's seen so far, it can rule out the 'ramp' explanation. While the 'agent' explanation is consistent with the data, it's less likely overall due to its lower prior probability. This implies that the model predicts that cubes will generally end up on the side that the ramp faces. The model also captures that participants made different selections depending on whether the manipulated feature was causally relevant (the cube color, Figure 14a) or not (the ramp color, Figure 14b). It does so by assuming more uncertainty in the ramp friction than to cube friction. This means that when the model simulates what would happen if the gray cubes slide down the colored ramps (e.g., Figure 14b), the two cubes end up roughly in the same position on average – no matter which ramp they are placed on. In contrast, when the two colored cubes slide down the gray ramp (e.g., Figure 14a), the red one tends to go further than the black one because the model adds less noise to each cube's friction in the simulations. Notice that the pattern of generalization was largely symmetrical: participants

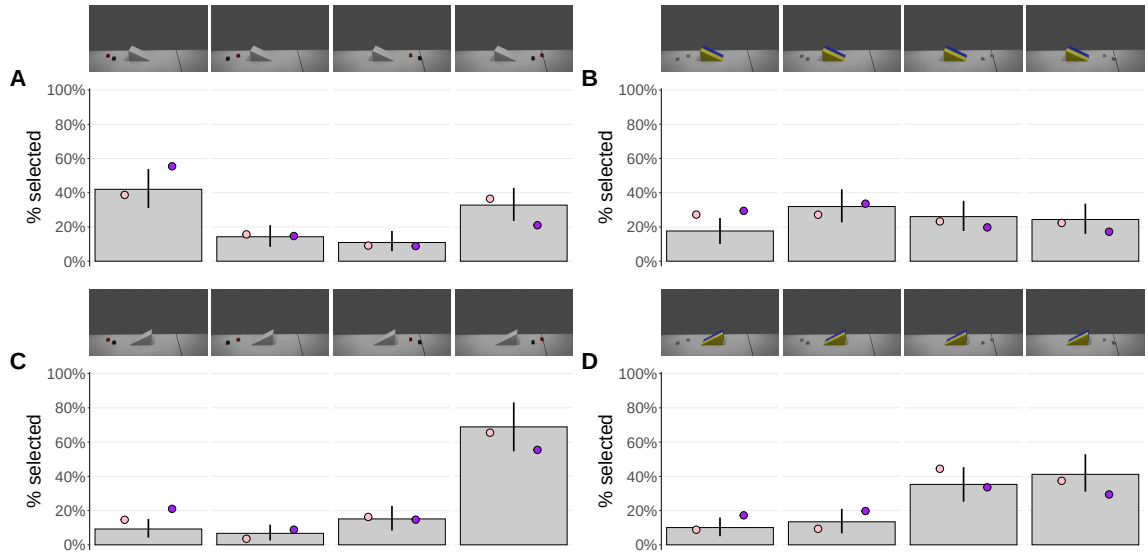

Figure 15

Experiment 3*. Participant selections in the 'generalization task' in the 'ramp backward' condition together with the predictions of the causal abstraction model (○) and the feature-based model (●). Note: Error bars are 95% bootstrapped confidence intervals. Notice that the results here are shown based on the combined data across all counterbalanced conditions, and illustrated for the case in which the cube featured mattered (rather than the ramp feature). Here, the black cube has more friction than the red cube, and the yellow ramp has more friction than the blue ramp.*

selections for the 'ramp backwards' trials (Figure 14c and Figure 14d) mirrored those of the 'ramp forward' trials (Figure 14a and Figure 14b).

The feature-based model captures this pattern of results by assuming that people pay more attention to the cube color than the ramp color (the `property` feature), that they learn which of the two colors is associated with the cube being further away from the ramp (the `friction` feature), and that people encode features in a relative manner, meaning that the final location of the cubes is encoded relative to the ramp orientation (the `side` feature). By assuming that these different features can interact with one another, the feature-based model is able to predict the response pattern in this condition.

Figure 15 shows the selections from the 'ramp backward' condition (Figure 8b). These participants saw trials c) and d) before a) and b). Notice first how this pattern of results looks quite different from that in the 'ramp forward' condition (Figure 14). Now, the pattern of responses to generalization trials when the ramp faces forward is no longer simply a mirror image of those when the ramp faces backward. Notice also, that the causal abstraction model still captures the overall pattern of responses, whereas the feature-based model struggles.

In Figure 15c and Figure 15d, participants thought that the cubes would end up on the right side of the backwards facing ramp, which is consistent with what they had experienced so far. In Figure 15c, most participants believed that the red cube would go

further, whereas in Figure 15d, a similar number of participants selected that one or the other cube would go further. The same idea applies again that people are more certain about the causally relevant feature (the cube color) than the irrelevant feature (the ramp color).

In Figure 15a, where the ramp now faced forward – something that these participants hadn't seen before – the distribution of responses was bimodal. Some participants thought that the cubes would end up on the left side of the ramp, whereas others thought that they would end up on the right side. Either way, most participants thought that the red cube would be further away from the ramp than the black cube. In Figure 15b, participants' responses were roughly uniform across the four options. Participants were equally likely to think that the cubes would end up on either side, and that either cube would be going further.

While the causal abstraction model uses the same prior beliefs for this experimental condition, this time, the data in the 'prediction task' rule out the 'physics' explanation. However, both the 'agent' and 'ramp' explanations are equally consistent with the data (Figure 11b). Both of these explanations agree about what happens if the ramp faces forward, but they make different predictions about what happens if the ramp faces backward. This is reflected in participants' bimodal distribution of selections in Figure 15a and in their uniform distribution in Figure 15b. Because the model's prior belief in the 'agent' and 'ramp' explanation is roughly equal, it correctly predicts that participants are just as likely to select images where the cubes are on the left or on the right, when the ramp faces forward in the generalization task.

Again, the model also captures the differences between the trials featuring unknown ramps versus cubes. The model has learned that the red cube has lower friction than the black cube, but it has more uncertainty about the friction of the two different ramps. So in Figure 15a, the model is uncertain whether the cubes end up on the right ('agent' explanation) or on the left ('ramp' explanation), but it does know that – either way – the red block would go further. In contrast, in Figure 15b, the model is uncertain about both the side where cubes would end up, and which one would go further.

Overall, the causal abstraction model accurately captures participants' selections in the generalization task. Across the two conditions and four generalization trials, it achieves a very good fit with $r = .99$, RMSE = 3.71 (the model predicts 32 different data points shown in Figure 14 and Figure 15 using 7 free parameters). In contrast, the feature-based model captures participants' selections less well than the causal abstraction model with $r = .96$, RMSE = 7.37. This is despite the fact that this model has 16 parameters in total compared to the 7 parameters of the causal abstraction model. Importantly, the featured-based model fails to capture that participants selections' in the 'ramp backward' condition are asymmetric in the generalization task.

We had pre-registered the hypothesis that the proportion of participants indicating that the cubes would be on the left or the right of the ramp would depend on how the ramp faced in their prediction task, and on how the ramp faced in the generalization task. Figure 16 shows these data. We ran a Bayesian logistic mixed effects model with ramp orientation in the training and on the generalization task as well as the interaction as fixed effects, and participant intercepts as random effects. We used sum contrasts to code the predictors ('training forward' = 1, 'training backward' = −1; 'generalization forward' = 1,

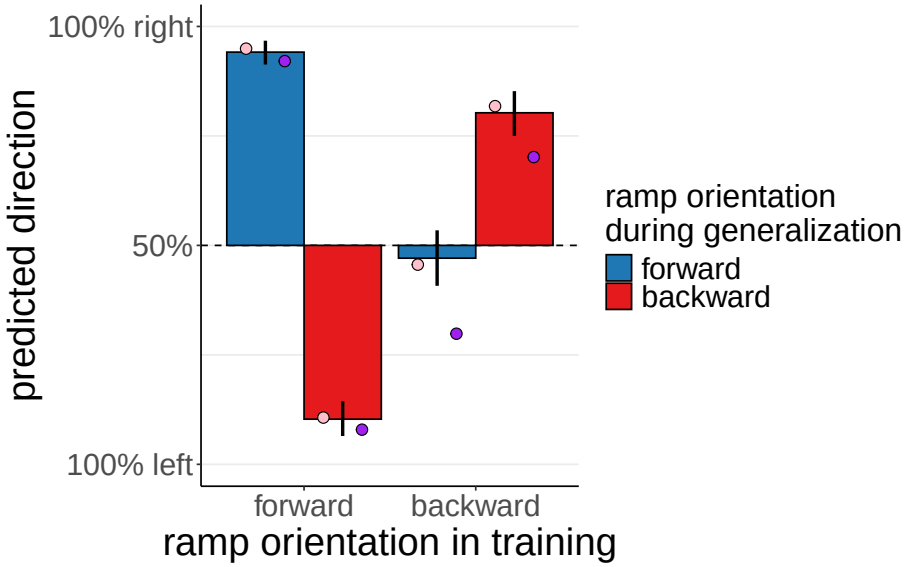

Figure 16

Experiment 3*. Participants' predictions of the direction in which the cubes would go as a function of how the ramp was oriented during training (x-axis) and during the generalization task (bar colors). Bars show the percentage of participants who predicted that the cubes would end up on the left or the right. For example, the left-most bar aggregates the selections in Figure 14a and Figure 14b, where most participants thought that the cubes would end up on the right side. Points show predictions by the causal abstraction model (●) and the feature-based model (●).* Note*: Error bars show 95% bootstrapped confidence intervals.*

'generalization backward' $= -1$). As predicted, we found a credible effect of generalization direction, $\beta = 0.90,\ 95\%\ CrI = [0.69, 1.12]$, and an interaction effect, $\beta = 1.70,\ 95\%\ CrI = [1.47, 1.95]$. While the effect of training was in the predicted direction, it wasn't credible, $\beta = -0.17,\ 95\%\ CrI = [-0.39, 0.04]$. Figure 16 also highlights what we mentioned before in terms of model predictions: while the causal abstraction model captures the asymmetric way in which participants generalize based on whether they experienced a forward or backward facing ramp in training, the feature-based model fails to do so. It predicts that within each condition, there would be a symmetric effect in the generalization trials.

Table 2

Experiment 3*. Number of participants whose selections were consistent with the predictions of the three different models ('physics', 'agent', or 'ramp'), or not fully consistent with either of the models ('other').*

ramp orientation in training	physics	agent	ramp	other
forward	95	6	0	19
backward	7	27	45	40

The selections in Figure 14, Figure 15, and Figure 16 are aggregated across all participants. We also looked at the response profiles of individual participants. We classified participants into four different groups based on what strategy they used (see Figure 11): 'physics', 'agent', 'ramp', and 'other'. A participant was classified as having used the 'physics' explanation when they answered that cubes would end up on the right side when the ramp was facing forward, and on the left when it was facing backward. A participant received the 'agent' label when they answered that cubes would end up on the right side no matter which way the ramp was facing. A participant received the 'ramp' label when they answered that the cube would end up on the left for forward facing ramps, and on the right for backward facing ramps. Finally, we classified a participant as 'other' when they made at least one selection that was not captured by any of the three explanations.

Table 2 shows the results of this analysis. Most participants' response profiles in the 'ramp forward' condition were consistent with the 'physics' explanation. In the 'ramp backward' condition, most participants' responses were consistent with the 'ramp' explanation, and many participants responded consistently with the 'agent' explanation. There was a larger proportion of 'other' participants whose responses weren't fully consistent with either of the three explanations in the 'ramp backward' condition compared to the 'ramp forward' condition.

Discussion

The results of the 'prediction task' and the 'surprise task' in Experiment 3 largely replicated what we had found in Experiment 2. Participants had no trouble learning to predict whether a cube would end up on the right side of the finish line even when they weren't shown any physical animations. It also didn't matter for their predictive accuracy whether the ramp faced forward or backward. Again, participants were likely to make mistakes in the 'surprise task' in a way that reflected that they preferentially paid attention to the feature that mattered for the prediction task (i.e., the color of the cube, or the color of the ramp). The pattern of participants' responses in the 'surprise task' was strikingly similar between the 'ramp forward' and the 'ramp backward' conditions.

So far, these results are consistent with the idea that people learn causal abstractions but are also consistent with a feature-based account where linear mapping between initial and final states are learned. In order to tease these two accounts apart, we added the 'generalization task' in Experiment 3. The causal abstraction model assumes that people come up with a causal story of how the data was generated. Participants in the 'ramp forward' condition are likely to infer that the cubes slide down the ramps (even though they never get to see any animations). Participants in the 'ramp backward' condition are likely to infer that cubes are pushed up the ramps. If participants indeed come up with different causal stories of how the data was generated, rather than merely learning a feature-based mapping between initial and final positions, then this should affect their predictions in the 'generalization task' in systematic ways. That's what we found. Participants' responses in the 'generalization task' were better predicted by the causal abstraction model than the feature-based model.

The feature-based model cannot explain why participants' responses differed systematically between the 'ramp forward' and 'ramp backward' conditions. When the orientation of the ramp changes in the generalization trial for the ramp backward condition, partici-

pants' selections are not simply a mirror image anymore. This is because, in this condition, participants came up with different explanations of what happened (as captured by the causal abstraction model), and these explanations are reflected in the generalization task.

General discussion

When people are asked to predict what will happen next, what do they learn about the situation? Across three experiments, we find evidence that people form abstract causal representations that are tailored to the specific task at hand. These causal abstractions are not simply based on associating scene-features with outcomes, but rather embody a causal story about how the observed data was generated. Crucially, the nature of this inferred causal story systematically influences how individuals generalize their knowledge to novel situations, in a way that purely feature-based models cannot explain.

In Experiment 1, participants learned which cubes turned on a blicket detector and which ones didn't. When only one feature mattered, such as the cube shape or color, participants learned this quickly. When we then surprised them and asked them what scene they had just seen, they remembered the diagnostic feature (say, the cube's shape), but they didn't remember the non-diagnostic feature (say, its color). Participants made systematic errors, revealing that they had learned a simplified representation that was sufficient for performing the task. Importantly, the recall errors only happened when participants had ample experience with the task. When participants had too little experience to form the relevant causal abstractions, they performed much better on the surprise test. A downside of this paradigm was that we only got one surprise task trial out of each participant, and that we couldn't rule out the possibility that participants remembered their final response to inform their reconstruction of what happened.

In Experiment 2, participants learned whether a cube that was placed on a ramp would cross a finish line. The cubes and ramps differed in their color. One of the features mattered for whether the cube would cross the finish line, while the other one didn't. However, both features together determined where exactly a cube ended up. Participants learned quickly whether a cube would end up crossing the finish line. This time, we surprised them by asking where exactly the cube would end up for each combination of cube and ramp color. Participants again made systematic errors. They were able to tell whether a cube would cross the finish line, but not where it would end up exactly.

Experiment 3 used the same paradigm as Experiment 2. However, this time, participants didn't watch animations of the cubes sliding down the ramp in the prediction task. Instead, they only saw the initial and final state. The experiment had two conditions: the ramp was either facing forward or backward. For both conditions, the cube in the final scene was shown to the right of the ramp. Participants performed similarly in the prediction task in both conditions, and they made similar errors in the surprise task – again, knowing which side of the finish line the cube would end up on, but not its exact position. In the generalization task, participants saw images of the two cubes placed on a novel ramp, or a novel cube placed on the two ramps that they had seen before. The ramp either faced forward or backward. Participants' responses in the generalization task were better predicted by a causal abstraction model than a feature-based model. The results are consistent with the idea that participants constructed different causal stories of what happened in the

two conditions. What story a participant inferred about the data determined how they generalized to new situations.

Our work builds on prior research showing that people learn and reason using abstract causal models rather than detailed, low-level representations of the world (Beckers & Halpern, 2019; Griffiths & Tenenbaum, 2009; Kelley, 1972; Schank & Abelson, 1977; Sloman, 2005). Complementing this, a separate line of research highlights that goals profoundly shape mental representations, influencing attention, learning, and memory to ensure efficient allocation of limited cognitive resources by prioritizing goal-relevant information (Bates et al., 2019; Ho et al., 2022; Kaplan et al., 2017; Leong et al., 2017; Maruff et al., 1999). Our work here integrates these insights by investigating how individuals' goals guide the specific ways in which they abstract causal information.

Limitations and future directions

We studied causal abstraction in a small set of situations involving blicket detectors and cubes sliding down ramps. Future work needs to study the phenomenon of causal abstraction more broadly. Abstraction can take many forms (Ibeling & Icard, 2023; Son, Vives, Bhandari, & FeldmanHall, 2024; Tenenbaum et al., 2011). Here, we focused on a particular kind of abstraction: compression. Participants learned representations that abstracted away some of the details of what actually happened. Future work should explore other aspects of abstraction such as how people learn hierarchical representations. The causal abstraction model stipulated different stories of how the data may have been generated (see Figure 11). More work is needed to figure out how people construct different candidate stories in the first place (for example, it's plausible that we begin assuming a simple physical story and only postulate additional factors if needed – rather than consider the three hypotheses from the beginning; see Bramley, Lagnado, & Speekenbrink, 2015; Zhao, Lucas, & Bramley, 2024). Relatedly, the causal abstraction model captures whether people construe different variables in a model more finely or coarsely, but it doesn't address how people choose what variables to include in their mental model. We assumed that participants learned what variables to pay attention to over the course of the 'prediction trials'. However, we didn't model this learning phase explicitly. Finally, there are limits to what behavioral tests can reveal about people's mental models. Future work may explore neuro-scientific evidence, too. This may help, for example, to tease apart whether causal abstraction happens primarily during the encoding stage or decoding stage of information processing.

Conclusion

Abstraction is critical for human thought. What we represent about the world, and how we represent it, depends on our goals: people build simplified representations that contain just what they need to achieve their goal. Importantly, this representation often includes a causal story of how the data was generated. What causal story people tell themselves about the data determines how they generalize to new situations.

Acknowledgments

TG was supported by research grants from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and Cooperative AI. We thank Thomas Icard and David Rose for helpful feedback. We thank Sarah Wu and Veronica Boyce for conducting reproducibility checks. Data from Experiments 1 and 2 have appeared in Shin & Gerstenberg (2023). Learning what matters: Causal abstraction in human inference. *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

References

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, *113*(27), 7345–7352.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2022). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, 1–12.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2), 1–23.

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.

Beckers, S., Eberhardt, F., & Halpern, J. Y. (2020). Approximate causal abstractions. In *Uncertainty in artificial intelligence* (pp. 606–615).

Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 2678–2685).

Beller, A., & Gerstenberg, T. (2025). Causation, meaning, and communication. *Psychological Review*.

Bigelow, E. J., McCoy, J. P., & Ullman, T. D. (2023). Non-commitment in mental imagery. *Cognition*, *238*, 105498.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109.

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

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Chalupka, K., Eberhardt, F., & Perona, P. (2017). Causal feature learning: an overview. *Behaviormetrika*, *44*(1), 137–164.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, *37*(6), 1171–1191.

Craik, K. J. W. (1943). *The nature of explanation*. Cambridge, UK: Cambridge University Press.

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72. Retrieved from `https://doi.org/10.1016%2Fj.artint.2015.12.003`

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258–1270.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity.

American psychologist, *52*(1), 45.

Gershman, S. J. (2018, May). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28. Retrieved from http://dx.doi.org/10.3758/s13423-018-1488-8 doi: 10.3758/s13423-018-1488-8

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*(6), 936–975.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Lawrence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–653). MIT Press.

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

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

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136.

Ho, M. K., Abel, D., Griffiths, T. L., & Littman, M. L. (2019). The value of abstraction. *Current Opinion in Behavioral Sciences*, *29*, 111–116.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: a theoretical integration with bayesian causal models. *Journal of Experimental Psychology: General*, *139*(4), 702–727.

Ibeling, D., & Icard, T. (2023). Comparing causal frameworks: Potential outcomes, structural models, graphs, and abstractions. *Advances in Neural Information Processing Systems*, *36*, 80130–80141.

Iwasaki, Y., & Simon, H. A. (1994). Causality and model abstraction. *Artificial intelligence*, *67*(1), 143–194.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Kaplan, R., Schuck, N. W., & Doeller, C. F. (2017). The role of mental maps in decision-making. *Trends in Neurosciences*, *40*(5), 256–259.

Kelley, H. H. (1972). *Causal schemata and the attribution process.* New York: General Learning Press.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107–128.

Kinney, D., & Lombrozo, T. (2024a). Building compressed causal models of the world. *Cognitive Psychology*, *155*, 101682.

Kinney, D., & Lombrozo, T. (2024b). Tell me your (cognitive) budget, and i'll tell you what you value. *Cognition*, *247*, 105782.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759. Retrieved from https://doi.org/10.1016%2Fj.tics.2017.06.002 doi: 10.1016/

j.tics.2017.06.002

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, 1–72. Retrieved from `http://dx.doi.org/10.1017/s0140525x16001837`

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*(2), 451–463.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2016). A bayesian theory of sequential causal learning and abstract transfer. *Cognitive Science*, *40*(2), 404–439.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208.

Maruff, P., Danckert, J., Camplin, G., & Currie, J. (1999). Behavioral goals constrain the selection of visual information. *Psychological Science*, *10*(6), 522–525.

Muhle-Karbe, P. S., Sheahan, H., Pezzulo, G., Spiers, H. J., Chien, S., Schuck, N. W., & Summerfield, C. (2023). Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *bioRxiv*.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* Hillsdale, NJ: Erlbaum.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

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

Smith, K. A., Hamrick, J. B., Sanborn, A. N., Battaglia, P. W., Gerstenberg, T., Ullman, T. D., & Tenenbaum, J. B. (2024). Probabilistic models of physical reasoning. In T. L. Griffiths, N. Chater, & J. B. Tenenbaum (Eds.), *Reverse engineering the mind: Probabilistic models of cognition.* MIT Press.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, *42*(6), 1103–1115.

Son, J.-Y., Vives, M.-L., Bhandari, A., & FeldmanHall, O. (2024). Replay shapes abstract cognitive maps for efficient social navigation. *Nature Human Behaviour*, *8*(11), 2156–2167.

Sosa, F. A., Gershman, S. J., & Ullman, T. D. (2025). Blending simulation and abstraction for physical reasoning. *Cognition*, *254*, 105995.

Strelnikoff, S., Jammalamadaka, A., & Lu, T.-C. (2022). Semantic causal abstraction for event prediction. In *International cross-domain conference for machine learning and knowledge extraction* (pp. 188–200).

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296. Retrieved from `http://dx.doi.org/10.1111/cogs.12605` doi: 10.1111/cogs.12605

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216–227.

Wellen, S., & Danks, D. (2014). Learning with a purpose: The influence of goals. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Williams, B. C. (1991). Critical abstraction: Generating simplest models for causal explanation. In *Proceedings of the fifth international workshop on qualitative reasoning about physical systems.*

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50.

Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology.* Oxford University Press.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28, pp. 127–135). MIT Press.

Yablo, S. (1997). Wide causation. *Philosophical Perspectives*, *11*, 251–281.

Zennaro, F. M., Turrini, P., & Damoulas, T. (2022). Towards computing an optimal abstraction for structural causal models. *arXiv preprint arXiv:2208.00894*.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2024). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, *8*(1), 125–136.