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

September 2, 2026

Author Note

[*]Corresponding author: Tobias Gerstenberg

https://orcid.org/0000-0002-9162-0779 (gerstenberg@stanford.edu)

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 more consistent with the causal account than 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.

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

Public Significance Statement

People cannot keep track of every detail of their surroundings, so they build simplified mental models that capture whatever matters for the task at hand. Across three experiments, we show that people's mental models can take the form of abstract causal stories, not just simple associations between features and outcomes: participants systematically forget or misremember details that were irrelevant to their goal, and the causal story they settle on to explain what they saw shapes how they apply that knowledge to new situations. These findings help explain how people manage to act intelligently in a complex world despite representing only a small part of it, and may inform how we design artificial systems that are both intelligent and efficient.

Introduction

One prominent view in cognitive science suggests that much of human learning can be understood as mental model building (Craik, 1943; Simon, 1979; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Accordingly, mental models reflect key aspects of the world and, through the process of mental simulation, can be used to make predictions about the future (Battaglia, Hamrick, & Tenenbaum, 2013; Smith & Vul, 2013), inferences about the past (Beller et al., 2025; Smith & Vul, 2014), and counterfactual judgments about how things could have turned out differently (Beller & Gerstenberg, 2026; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Zhou, Smith, Tenenbaum, & Gerstenberg, 2023).

In recent years, researchers have found that people's judgments on tasks that require intuitive physical understanding can be well explained by the idea that their mental models function similarly to physics engines that create realistic animations in modern computer games (Kubricht, Holyoak, & Lu, 2017; Smith et al., 2024; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). This idea has been used to explain how people judge whether a block tower will fall (Battaglia et al., 2013), whether one billiard ball caused another to go through a gate (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017), or how far a block will slide down a ramp (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). In this approach, people's physical judgments are modeled by taking the physics engine that created the stimuli as a starting point, and then injecting noise into the simulator to capture people's uncertainty. For example, to match people's graded judgments about a tower's stability, the model assumes uncertainty about the exact position of each block, and then simulates what would happen if the block positions were slightly perturbed. And to capture people's graded causal judgments about collision events, the model assumes uncertainty about how balls move, and then simulates what would have happened if the candidate cause hadn't been present in the scene.

One limitation of this approach is that it requires a full specification of the physical scene. Each object needs to have an exact size, location, mass, friction, elasticity, etc.

However, people's mental models are unlikely to represent the physical world at this level of detail (Bigelow, McCoy, & Ullman, 2023; Davis & Marcus, 2016). For instance, people may not simulate the whole scene but only run partial simulations instead (Bass, Smith, Bonawitz, & Ullman, 2022). Also, while many things happen in parallel in the physical world, people often simulate events sequentially in their mind (Balaban & Ullman, 2025). People have limited cognitive resources, so they cannot represent everything, everywhere, all at once (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?

Abstract causal models

Abstract schemas and scripts organize actions and structure memories of what happened (Kelley, 1972, 1973; Schank & Abelson, 1977). Often, 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 to new situations (Bareinboim & Pearl, 2016; Lake, Ullman, Tenenbaum, & Gershman, 2017). Causal models can be formulated at many levels of abstraction (Griffiths & Tenenbaum, 2009; Iwasaki & Simon, 1994). For example, we might represent the act of opening a door with a simple binary cause (a door knob that's either turned or not) and a simple binary effect (a door that's either open or not). This level of abstraction is helpful for formulating a higher-level plan: I need to turn the knob to open the door. Ultimately, however, opening a door requires a detailed execution of a motor plan in a high-dimensional action space that's continuous in space and time. What level of abstraction is right depends on the task at hand – in this example, whether that task is planning an action versus taking it (e.g., Beckers, Eberhardt, & Halpern, 2020; Beckers & Halpern, 2019; Chalupka, Eberhardt, & Perona, 2017; Strelnikoff, Jammalamadaka, & Lu, 2022; Yildirim, Gerstenberg, Saeed, Toussant, &

Tenenbaum, 2017).

The question of how to choose the right level of abstraction has been studied extensively in philosophy, cognitive science, and computer science. One proposal from philosophy suggests that the variables in a causal model should be 'proportional' to one another (Woodward, 2021; Yablo, 1997). Accordingly, cause and effect variables should be specified at a level of detail that matches. For example, binary variables work well for representing a light switch and a light bulb. To represent a dimmer and a dimmable light bulb, continuous variables work well. But, pairing a continuous variable (e.g., a dimmer) with a binary one (e.g., a light bulb that can only ever be on or off) wouldn't be proportional.[1]

In the literature on causal learning, there is a helpful distinction between learning the causal structure (e.g., Bramley, Dayan, Griffiths, & Lagnado, 2017; Meder, Mayrhofer, & Waldmann, 2014; Waldmann, Holyoak, & Fratianne, 1995) and learning the causal strength between variables (e.g., Cheng & Novick, 1991; Griffiths & Tenenbaum, 2005; Holyoak & Cheng, 2011). Learning the causal structure means figuring out how different variables are causally related, and how multiple variables combine to influence a common effect. Learning the causal strength means figuring out how strongly one variable affects another. This distinction highlights that it's possible for a learner to have a good sense of what causes what (i.e., the causal structure), without knowing exactly how likely it is that one thing will make another thing happen (i.e., the causal strength). To learn how a causal system works, people draw on a variety of information including prior knowledge (Griffiths & Tenenbaum, 2009; Waldmann & Hagmayer, 2001), statistical co-variation between events (Cheng & Novick, 1992; Shanks & Dickinson, 1987), temporal contiguity (Bramley,

[1]While proportionality is a useful feature in general, there are of course many instances where variables with different levels of granularity are paired well together. For example, one variable might represent the number of votes for one candidate, and another variable in the same model might represent which candidate won.

Gerstenberg, Mayrhofer, & Lagnado, 2018; Lagnado & Sloman, 2006), and explanations from others (Keil, 2006; Lombrozo & Vasilyeva, 2017).

Past research has shown that people are good at learning abstract causal models that capture the essential structure of a system (Wellen & Danks, 2014). For example, adults (Lu, Rojas, Beckers, & 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. Here, the learning goes beyond that of any specific cause-effect relationship to more general causal structures that may also hold true for other causal systems (Griffiths & Tenenbaum, 2009; Schulz, Goodman, Tenenbaum, & Jenkins, 2008). Even though 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. They constrain what we attend to (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017; Maruff, Danckert, Camplin, & Currie, 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 (Altmann & Trafton, 2002; Ho et al., 2022), and what we believe (S. J. Gershman, 2018; Kunda, 1990). Because people's cognitive resources are limited, they need to allocate them efficiently (Bates, Lerch, Sims, & Jacobs, 2019; Belledonne, Butkus, Scholl, & Yildirim, 2026; 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, they remember its location well when it affected their planned route, but less well when it didn't (Ho et al., 2022; see also Chen, Cheyette, Allen, Tenenbaum, & Smith, 2026). 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 (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.

Attention in learning

The influence of goals on cognitive processing has been studied extensively in the context of attention (e.g., Mackintosh, 1975; Pearce & Hall, 1980). According to one view, people attend to *predictive* cues – cues that are good predictors of the outcome of interest (Mackintosh, 1975). On another view, people attend to *surprising* cues – cues that they are highly uncertain about (Pearce & Hall, 1980). These two views paint different pictures of how attention works; whether it's primarily goal-driven or stimulus-driven (see Holland & Schiffino, 2016; Le Pelley, Mitchell, Beesley, George, & Wills, 2016, for reviews).

One common way in which the role of attention in learning is studied is via the selective-attention category-learning task (Shepard, Hovland, & Jenkins, 1961). In this task, participants have to learn how to categorize various stimuli, such as images of geometric shapes that differ in shape, color, and size. In some versions of the task, the relevant stimulus dimension changes over time such that, for example, shape matters first but then color (e.g., Leong et al., 2017; Niv et al., 2015; Unger & Sloutsky, 2023). The general finding is that participants quickly learn what aspect of the stimulus is relevant (e.g., the shape) and what aspect doesn't matter (e.g., the color). Over the years, several computational models have been developed that capture participants' learning in tasks like these (Kruschke, 1996, 2001; Love, Medin, & Gureckis, 2004; Nosofsky, 1986). For example, Leong et al. (2017) developed an attention-weighted model that learns to place more weight on the relevant stimulus dimensions and ignore the irrelevant ones, thereby implicitly constructing a simplified representation of the task (see also Belledonne et al., 2026; Mack et al., 2016; Mack, Preston, & Love, 2020).

Representation learning

The question of how to construct task representations that support efficient learning, planning, and decision-making is also at the heart of reinforcement learning (Niv, 2019; Sutton & Barto, 1998). In reinforcement learning, agents take actions that move them from one state to another, following a transition function that encodes the causal structure of the world. Once the state space and action space have been defined, an agent's learning objective is to find an optimal policy for taking (costly) actions that yield (rewarding) states. Now, the problem is that, in the real world, agents aren't handed a convenient state and action representation. Generally, the environment is high-dimensional but only some of this information is causally relevant for the task at hand. This means that agents have to learn what information to ignore (*selective attention*; S. Gershman, Cohen, & Niv, 2010; Niv et al., 2015; Wilson & Niv, 2012), and what the observable information reveals about the latent, task-relevant causal structure (*causal inference*; S. J. Gershman, Norman, & Niv, 2015). Because no two situations are ever fully alike, learning the right state abstractions is critical for generalizing what was learned from one situation to another without over-generalizing to superficially similar situations (Collins & Frank, 2013; Xia & Collins, 2021). Similarly to the state space, the action space can be very high-dimensional, too, and evolve continuously in space and time, such as when a robot wants to dynamically interact with a physical environment (Konidaris, 2019). Generally, state and action representations constrain one another. For example, if an agent can only ever take two actions, then they don't need a very rich state representation (which is related to the idea of 'proportionality' that we mentioned earlier; Woodward, 2021).

From a computational perspective, there are two main approaches for representation learning: deep learning and program induction. The deep learning approach uses large artificial neural networks that discover relevant features through supervised or unsupervised learning. These networks often learn to represent more superficial features in earlier layers, and more abstract features in deeper layers (Bengio, Courville, & Vincent,

2013; Mienye & Swart, 2025; Mnih et al., 2015). The program induction approach uses symbolic structures as the representation language and aims to infer what computer program generated the observed data (Lake et al., 2017; Tenenbaum et al., 2011). Recent work attempts to combine the best of both worlds by relying on general, learned associations to build specific, structured representations that are tailored to the task at hand (Wong et al., 2025). The problem of representation learning is a key area of interest in neuroscience, too, with some evidence that both kinds of learning mechanisms are implemented in the brain (Muhle-Karbe et al., 2023; Schuck, Cai, Wilson, & Niv, 2016; Schuck et al., 2015; Schuck & Niv, 2019; Sharpe et al., 2019; Tian et al., 2026).

Goal-dependent abstract causal models

Much work on representation learning in psychology and neuroscience has used relatively simple tasks to study abstraction, such as the selective-attention category-learning task (Niv, 2019; Shepard et al., 1961). In such tasks, the successful learning often consists of finding a good direct mapping from features to outcomes. Here, we are interested in how people build *abstract causal models* of the physical world, where these models postulate hidden properties that capture the causal dynamics of how the observable data is produced. When building causal models of the world, the learner has to know what information to represent (Icard III & Goodman, 2015), and how.

Most work on abstraction in causal models has been theoretical and not connected directly with how humans think (Beckers & Halpern, 2019; Chalupka et al., 2017; Geiger et al., 2025; Zennaro, Turrini, & Damoulas, 2022). However, recently, Kinney and Lombrozo (2024a) developed 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 coarse level of granularity (see also Pacer & Lombrozo, 2017). However, simpler representations encode less information. How should one balance the trade-off between

compression and informativeness? Kinney and Lombrozo (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 mentioned above) and stability. A causal variable is *proportional* with respect to an effect variable when little to no information would be gained by replacing it 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) showed 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 also showed that people can infer what a person values from how detailed their causal explanation was (Kinney & Lombrozo, 2024b).

Our contributions

We know that people are capable of learning abstract causal models (Lu et al., 2016; Lucas et al., 2014; Wellen & Danks, 2014), and we know that people don't represent everything about a situation but focus on what's relevant for their goals (Ho et al., 2022; Niv et al., 2015). What we don't know yet is how these two capacities fit together. When someone learns to predict what will happen in an unfamiliar environment, do they build a

causal model of how the data was generated and, if so, what determines the level of detail at which they represent it?

Our paper makes the following contributions. First, we develop a computational account that captures how people learn an abstract causal model of a domain, and how they use that model to make predictions about scenes they haven't seen before. Whereas prior work has focused on learning problems that can be modeled with causal Bayesian networks, we extend this work to scenes in which physical objects dynamically interact in continuous space and time.

Second, we develop two novel paradigms for testing causal abstraction. Unlike prior work that has focused on relatively simple categorization tasks, our paradigms employ a distinctly causal structure. That way, we can test whether and how people construct mental representations that take the form of abstract causal models versus models that directly map from observable features to outcomes.

Third, by pairing these paradigms with memory, generalization, and explanation tasks, we can probe the cognitive consequences of constructing abstract causal models. Our tasks reveal the costs of abstraction: participants misremember features they didn't need earlier, and they are uncertain about what will happen in new situations where these features now matter. What our tasks don't show as directly are the benefits of abstraction – that abstracting made people learn faster, or transfer knowledge better.

Overall, our work brings together several literatures that have been fairly isolated so far: work on selective attention, work on causal abstraction, and work on physical reasoning. Work on selective attention has mostly used categorization tasks in which the learner's job is to map features onto outcomes (Leong et al., 2017; Niv et al., 2015; Shepard et al., 1961). Such tasks show that people learn to ignore what doesn't matter, but they can't tell us whether what people end up with is a causal model of the situation, or simply a mapping from predictive features to outcomes. Work on causal abstraction has been largely theoretical (Beckers & Halpern, 2019; Chalupka et al., 2017) or, where it has been

tested empirically, has asked people to evaluate causal descriptions supplied by the experimenter rather than to construct their own from experience (Kinney & Lombrozo, 2024a). Finally, work on physical reasoning has used models that rely on full scene specifications to capture people's judgments (although see Chen et al., 2026). Here, we propose and test a model of causal abstraction that takes in a physical scene, constructs a causal story of what happened, and fills in those details of the story that matter. By developing experimental paradigms that allow for novel ways of probing people's knowledge, we provide evidence that people learn abstract causal models that are tailored to the task at hand.

Overview of the experimental paradigms and approach

We investigate whether people learn and use goal-dependent abstract causal models across two paradigms: the 'blicket paradigm', and the 'ramp paradigm'. 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 et al., 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 an object 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 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

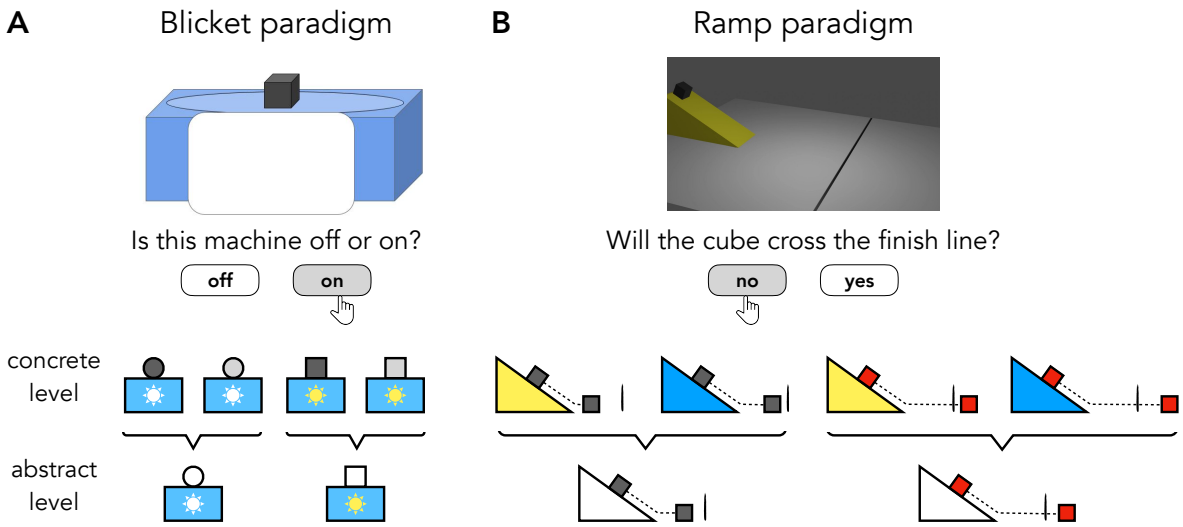

Figure 1

*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.*

what happens (just as 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 make systematic memory errors that are consistent with having learned goal-dependent abstractions; they are more likely to remember goal-relevant features than irrelevant ones. 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. While consistent with causal abstraction, the results of Experiments 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 a rich causal model). Experiment 3 creates a targeted situation where participants have to generalize what they have learned, and where the two accounts make different predictions. We find that the causal abstraction model better captures participants' judgments than the feature-based model. What abstract causal model participants learned determined how they generalized their knowledge to new situations.

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

Transparency and openness

We report how we determined our sample size, all manipulations, and all measures for each of the three experiments. There were no data exclusions in any of the experiments. 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 links to 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"). Any analyses not specified in a pre-registration are noted as exploratory in the results sections.

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*: $N = 120$).[2] 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.[3] 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 black cube, a white cube, a black cylinder, or a white 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, its color, or the combination of the two. When only one of the two features mattered, we counterbalanced which one it was 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.

---

[2]Demographic information in all experiments was collected at the end of the study, with a text box for indicating age, and forced choice options for gender ['Female', 'Male', 'Non-binary', 'Other'], race ['White', 'Black/African American', 'American Indian/Alaska Native', 'Asian', 'Native Hawaiian/Pacific Islander', 'Multiracial/Mixed', 'Other'], and ethnicity ['Hispanic', 'Non-Hispanic']. If a participant selected 'Other' for gender or race, they were able to write their answer in a text box. Responses to the demographic survey were optional, meaning that participants were able to finish the study without providing any or only partial demographic information.

[3]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.

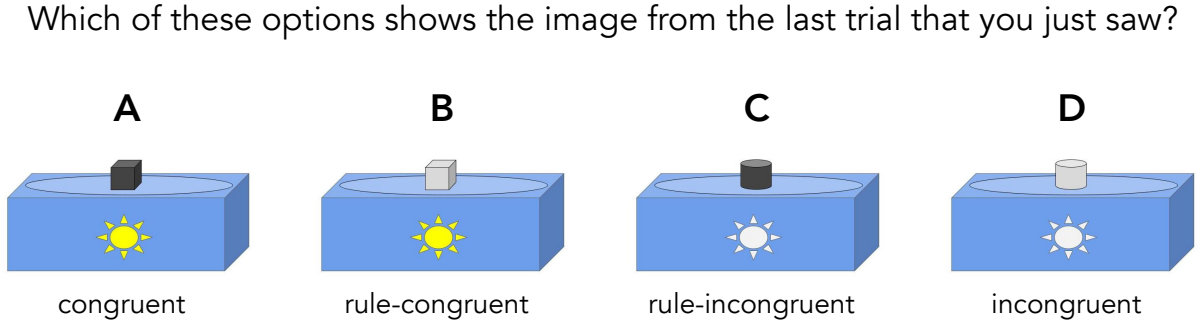

Figure 2

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).*

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 as 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 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 of four images they had just seen (see Figure 2). Finally, participants performed an explicit 'rule comprehension task' where they were shown each of the four objects on the same screen with a checkbox underneath each

one and asked "Which one of these blocks is a Blicket?" At the end of the experiment, participants provided demographic information.

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 both 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 white 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 partially matching responses (both the black cylinder and the white 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

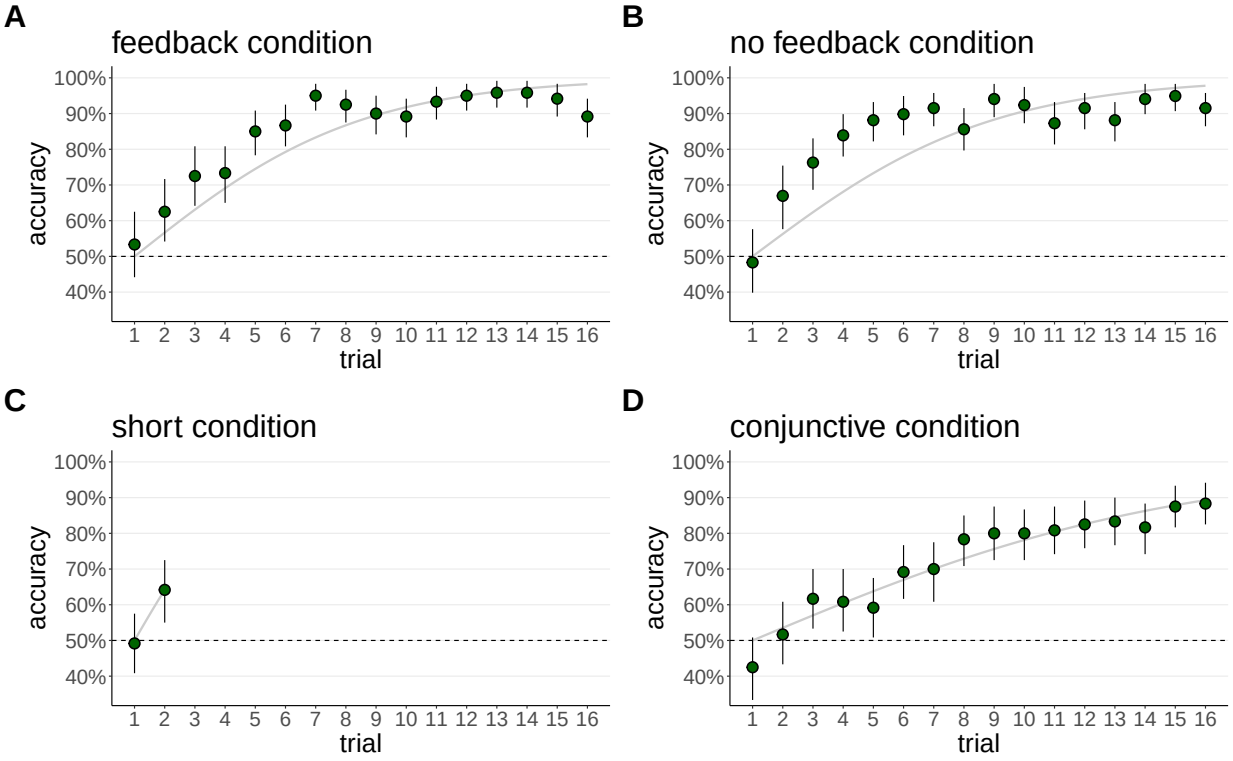

Figure 3

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.*

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

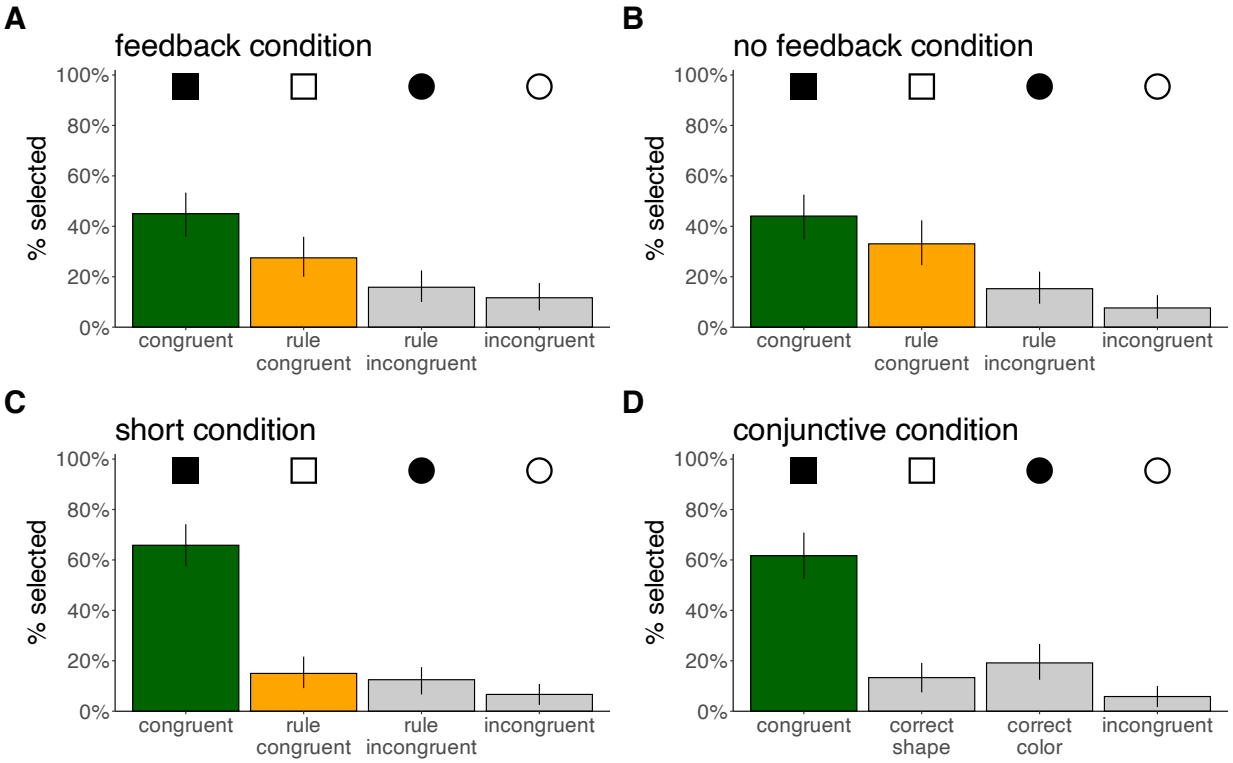

Figure 4

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.*

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, 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', the 'surprise task', and the 'rule comprehension' 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 would 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 $[0.21, 1.33]$, $p = .01$.[4]

In the 'short' condition we correctly predicted that there would be no significant difference in selecting the rule-congruent versus rule-incongruent option, $\beta = 0.18$,

[4]The pre-registrations for the four conditions of Experiment 1 were not fully consistent with one another in how the significance test was specified: the 'no feedback' condition pre-registered a one-tailed test ($\alpha = .05$), the 'short' condition pre-registered a two-tailed test, and the 'feedback' and 'conjunctive' conditions did not specify a directional test. We report all four conditions using the same two-tailed 95% CI and p-value convention for consistency; this does not change the pattern of significant versus non-significant results reported here.

95% CI $[-0.5, 0.87]$, $p = .6$.[5] 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.

Rule comprehension task

For each participant, we coded whether their overall responses in the 'rule comprehension task' were accurate or inaccurate. Note that in the conjunctive condition, only one out of the four objects was a blicket, while in the other three conditions, two out

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			
	feedback	no feedback	short	conjunctive
blicket	52%	47%	75%	77%
not a blicket	38%	42%	57%	57%

[5]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 $= -2.34, p = .05$.

of four objects were blickets. We counted a response as accurate only if they selected all of the correct blickets, and counted it as inaccurate otherwise. Participants' performance in the 'rule comprehension task' reflected their performance in the 'prediction task'. Their accuracy was high in the 'feedback' condition (80%), the 'no feedback' condition (82%), and the 'conjunctive' condition (72%), but low in the 'short' condition (12%).

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 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 option when asked what they had just seen in the 'feedback' and 'no feedback' conditions, 30% of participants selected the incorrect but rule-congruent option.

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 options. Put differently, when participants had ample time to learn (in the 'feedback' and 'no feedback' conditions), they performed markedly

worse on the 'surprise task'. 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 surprising 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.

One alternative explanation is that the effects we observed may in part be due to memory interference (e.g., Brown, Neath, & Chater, 2007). During the 'prediction task', participants saw multiple events that are consistent with the rule. So it's possible that they misremember an earlier (rule-congruent) instance, rather than the last one they had just seen.

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). As in Experiment 1, we manipulated two features: the color of the cube and the color of the ramp. Each feature was diagnostic of the object's friction. 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 to make predictions about where exactly the cube will end up (see Figure 5). By this time, participants will have seen what happens for each combination of ramp and cube, such that this 'surprise prediction task' can also be thought of as a 'surprise recall task'. 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 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. Our target sample size was 120 participants per condition, just as in Experiment 1. A total of 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. An object's color was diagnostic of 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

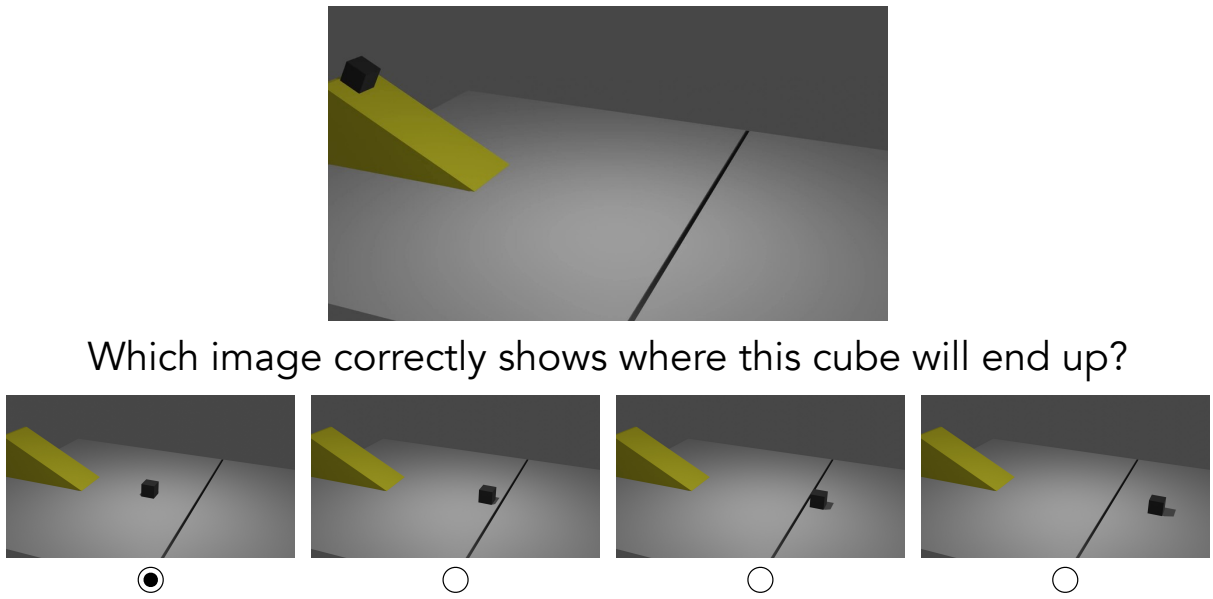

Figure 5

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.*

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 farther 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 prediction trials where participants were presented with an image showing a cube at the top of a ramp as 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?" as 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. At the end of the experiment, participants provided demographic information.

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 four trials (one for each combination of cube and ramp – presented in random order). 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 was causally relevant, as well as what colors were diagnostic of low and high friction.

Notice that the number of 'prediction task' trials in the 'short' condition here differs from that of Experiment 1, where participants only saw two trials. This is because the 'surprise task' differs between the two experiments. Whereas in Experiment 1, participants simply had to recall what had happened on the last trial, in Experiment 2, they need to predict (or recall) what happened in the four trials that combine each type of ramp with each type of cube. We wanted to make sure that participants in principle had all the information they needed to succeed at the 'surprise task'. If we had only shown participants two out of the four combinations of cubes and ramps, they would have needed to guess for the other two.

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.

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 select 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.

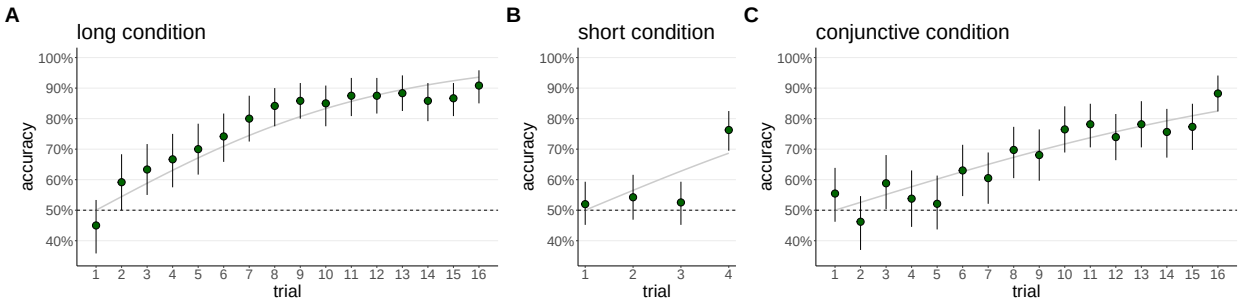

Figure 6

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.*

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. As 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 farther on blue than yellow ramps. The green bars in Figure 7 show correct responses, and the orange 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'

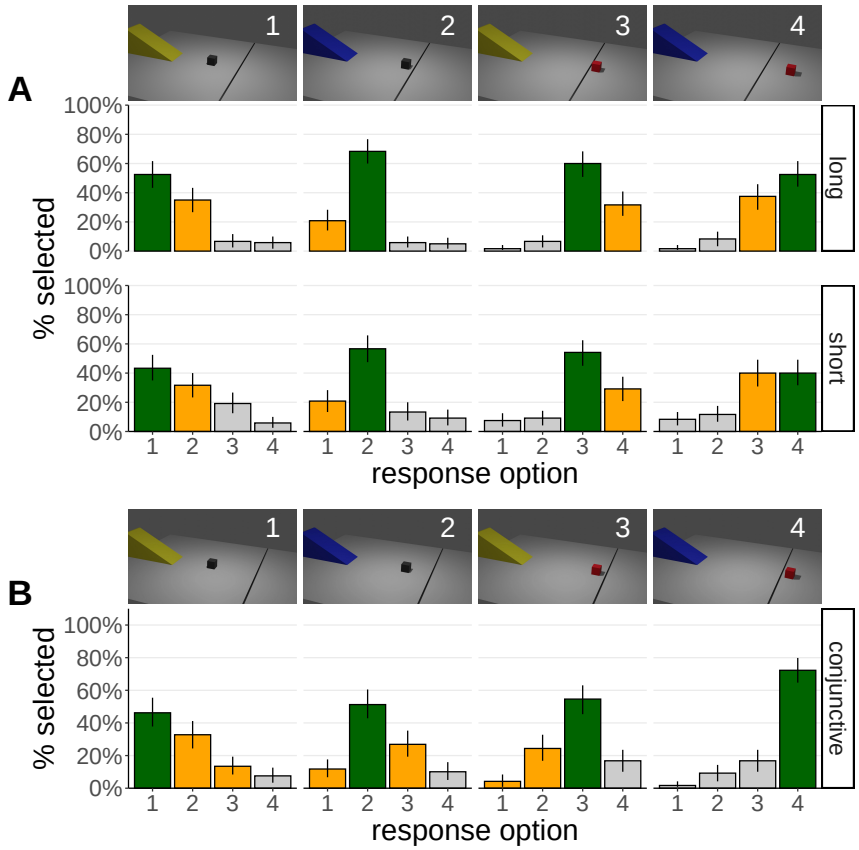

Figure 7

Experiment 2*. Participants' 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.*

selections 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.[6] 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

[6] 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).

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 shows a pattern of results that is again consistent with the idea that people learn abstract models that privilege causally relevant information. In the experiment, 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 had seen 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 quickly in this task. On the fourth trial, participants already had an accuracy of almost 80%. Remember 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 farther away ('long' condition: 52%; 'short' condition: 42%). In all three conditions, participants were very unlikely to select a response that was on the other side of the finish line from where the cube actually ended up. For example, participants were unlikely to select position 3 or 4 when the correct position was 1 or 2 (in the 'long' or 'short' conditions).

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 1 asked participants to recall something they had just seen. In Experiment 2, participants need to recall what happened in four different trials. Since we randomized the order of the prediction trials and of the four surprise trials, we can see whether participants performed better when the first surprise trial was of the same type as their last prediction trial – this renders the first surprise trial a little more similar to the 'surprise task' in Experiment 1. However, we found that participants were not more likely to be accurate in the first surprise trial when it matched the last trial in the 'prediction task' than when it didn't (see Table A1).

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 (non-causal) mapping from features to outcomes. For example, they may have 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 for a more critical test to better differentiate between the causal abstraction model and a feature-based model.

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' conditions, 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 that 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.

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.

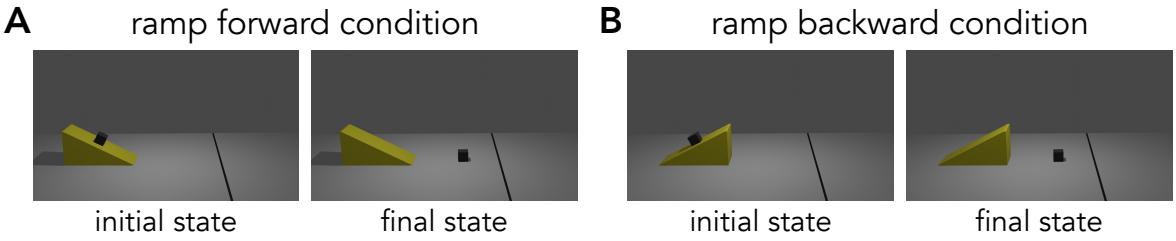

Figure 8

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.*

Finally, we also added an 'explanation task' where participants were asked to explain why they chose a particular response option in the 'generalization task'. We were curious to see whether this open-ended question would provide insights into how participants were thinking about the task.

Methods

Participants

A total of 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 4 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*: $N = 57$). Again, as in Experiments 1 and 2, our target sample size was 120 participants in each key experimental condition (here, across whether the ramp faced forward or backward).

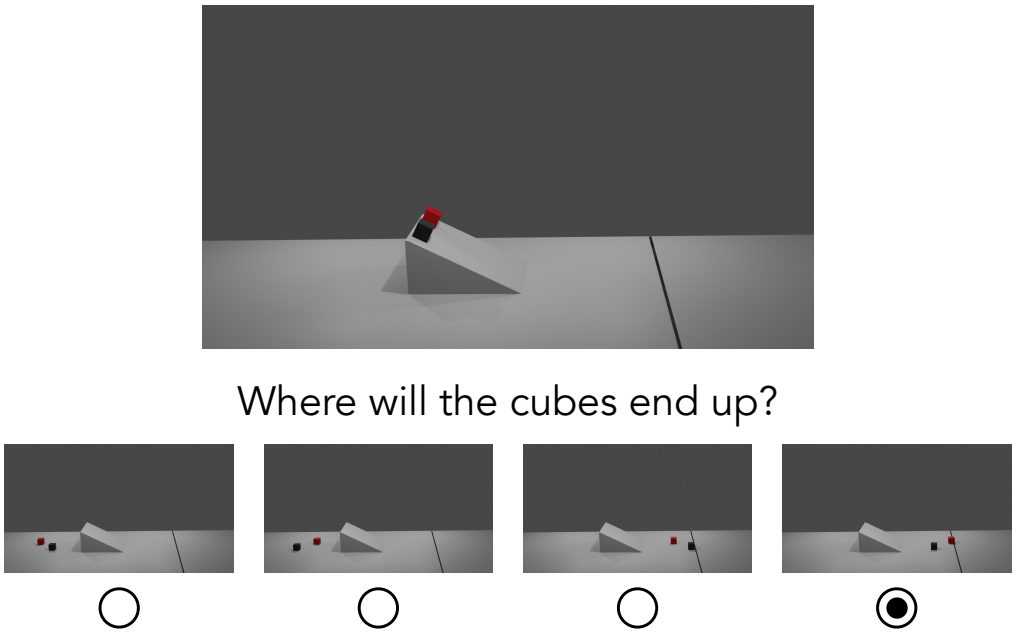

Figure 9

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.*

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 (four of each type of trial) in randomized order. The 'surprise task' was the same as in Experiment 2. Participants saw the four 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

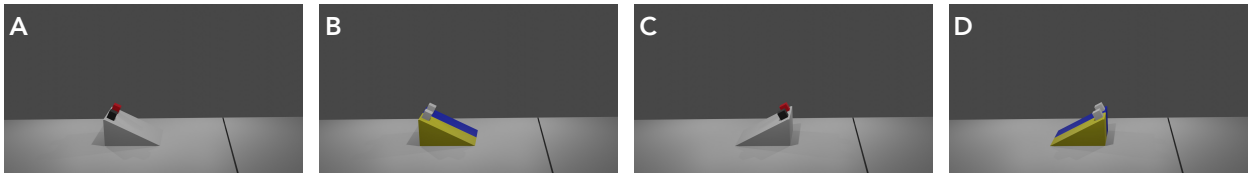

Figure 10

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 farther away from the ramp in one image, and the other cube in the other image.*

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 was shown in front for the generalization trials. For example, for some participants, the black cube was in front (as in Figure 9), and for others, the red cube. Finally, we added an 'explanation task' where participants saw what answer option they had selected for the four trials in the 'generalization task', one by one. For each one, we asked them: "For this image, you predicted that the cubes would end up as shown below. Why did you think this would happen?" Participants provided their answers in an open-ended text box.

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 the 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).

Predictions

We will discuss the predictions for the 'surprise task' and 'generalization task' in turn (we didn't pre-register any predictions for the 'explanation task').

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' conditions. 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.

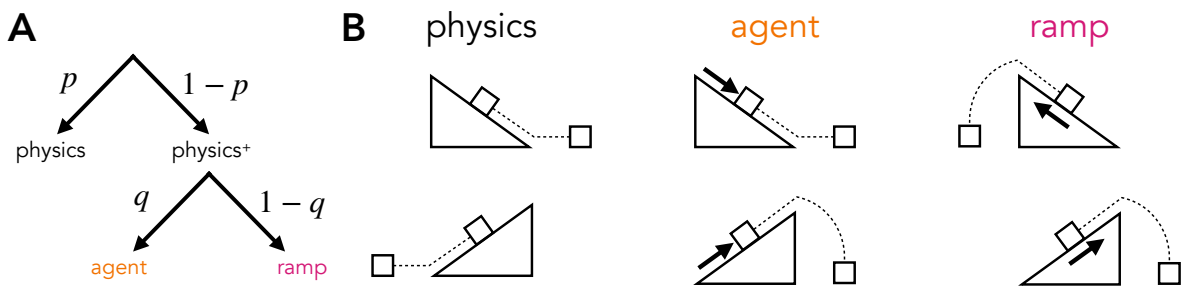

Figure 11

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 backward-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. The causal abstraction model operates in two steps: first, it generates a causal story of what happened (based on prior knowledge and initial evidence). Second, it then refines this causal story based on what information is relevant for the task at hand.

We construe participants' task of figuring out what happened and generalizing to new situations as a Bayesian inference problem.[7] Accordingly, the goal is to update one's prior beliefs about different hypotheses p(hypothesis) to a posterior belief p(hypothesis|data) by considering the likelihood of the observed data under each

[7]We provide a high-level description of the model here. For a more detailed discussion, see the appendix.

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 about 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 participants see in the 'ramp forward' condition is equally consistent with the 'physics' and the 'agent' hypotheses, but not with the 'ramp' hypothesis. Conversely, the data in the 'ramp backward' condition is equally consistent with the 'agent' and the 'ramp' hypotheses, but not with the 'physics' hypothesis.

Two free parameters in the model, p and q, characterize the model's prior belief over the different hypotheses (see Figure 11a). For example, we can assume that a priori, the model considers it more likely that the simple 'physics' hypothesis is true than the alternative hypotheses.

Based on the observed cube positions in the 'prediction task', the model updates its prior belief to a posterior over the different hypotheses. Here, we simply assume that the model (softly) rules out hypotheses that are inconsistent with the data. For example, if a cube is observed on the right side of a backward-facing ramp, both the 'agent' and 'ramp' hypotheses assign the same likelihood, whereas the 'physics' hypothesis assigns a very low likelihood. After the evidence from the 'prediction task', the model thus has formed a posterior over the different hypotheses that it then uses to make predictions about what will happen in the 'generalization task'.

To predict participants' selections in this task, the model computes the posterior predictive probability of the four response options

$$p(\text{option}_i|\text{data}) = \sum_{j \in \{\text{physics,agent,ramp}\}} p(\text{option}_i|\text{hypothesis}_j) \cdot p(\text{hypothesis}_j|\text{data}), \tag{2}$$

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 $p(\text{hypothesis}_j|\text{data})$ is the posterior probability of that hypothesis based on the data seen so far.

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. Importantly, we assume that participants have more uncertainty about the friction parameter that didn't matter for the 'prediction task'. So if, for example, it didn't matter what ramp a cube was placed on for whether it crossed the finish line, then the model is more uncertain (i.e., adds more Gaussian noise) about the ramp friction compared to the cube friction. To compute the posterior predictive probability for each of the four images under a given hypothesis, the model generates many noisy samples and then checks how likely the cubes would end up where the response options show them based on its distribution of samples. Finally, the model uses a softmax function to translate from the continuous posterior predictive probability into a discrete choice. Overall, this model has seven free parameters which we fitted by maximizing the likelihood of participants' generalization choices.

There is a close analogy between how the causal abstraction model works and learning causal structure and strength (e.g., Griffiths & Tenenbaum, 2005; Waldmann et al., 1995). The model considers a small set of possible causal structures that could have generated the data (the causal stories of what happened). For a given causal story, it then learns the causal strengths of the different features that affect the outcome (the ramp and the cube friction) based on how much they matter for the task at hand. If a feature doesn't matter, the model remains uncertain about it.

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 farther away (options 1 and 4). The 'inconsistent' options show the black cube farther 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:

$$p(\text{option}_i) = \texttt{1 + orientation * side * property * friction}, \tag{3}$$

where the 1 denotes the global intercept and the '`*`' indicates that in addition to each feature individually, their interactions are included in the regression model as well (including all two-way, three-way, and four-way interactions between the features). This model has 16 free parameters in total.

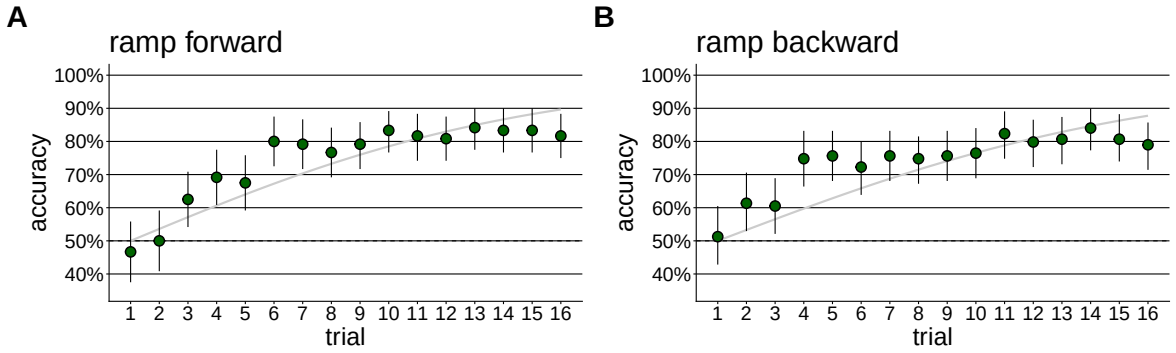

Figure 12

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.*

Results

We discuss the results of the 'prediction task', 'surprise task', 'generalization task', and 'explanation 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). 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 was 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.

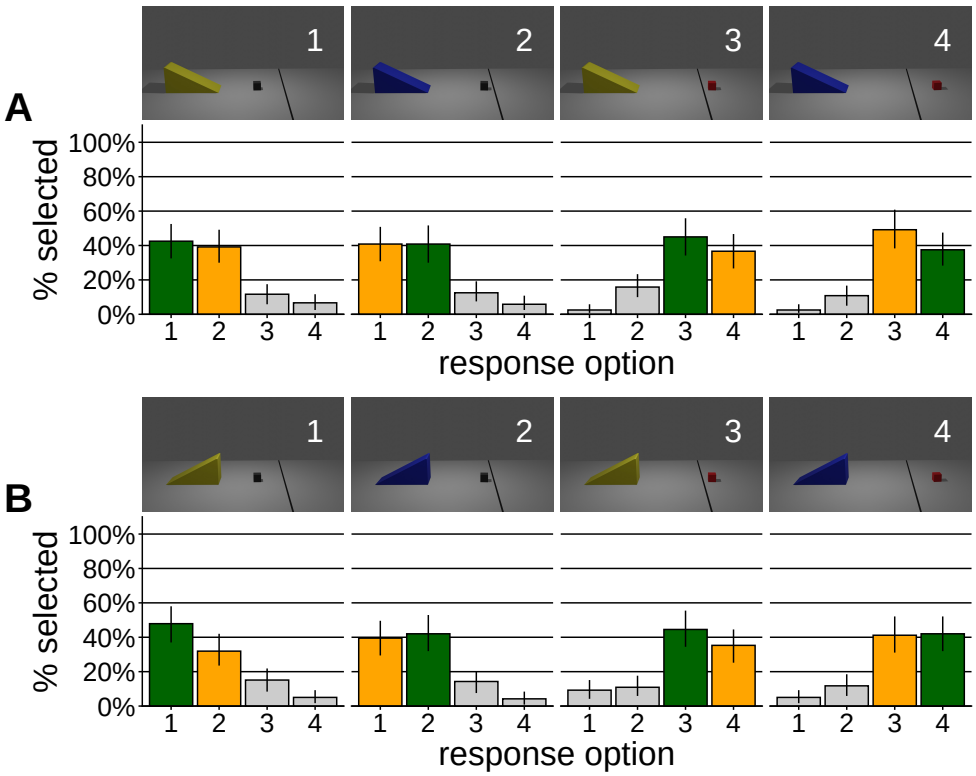

Figure 13

Experiment 3*. Participants' 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 as in Figure 7, we show the results aggregated across all counterbalanced conditions.*

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 (shown 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 farther away (positions 1 or 4).

Generalization task

The 'ramp forward' condition. Figure 14 shows the selections from the 'ramp forward' condition (Figure 8a). In Figure 14a, most participants thought that the cubes would end up on the right side of the ramp, and that the red cube would end up farther 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 farther 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 (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 farther 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 farther. Overall, the results in the 'generalization task' for the backward ramp closely mirror those for the forward ramp.

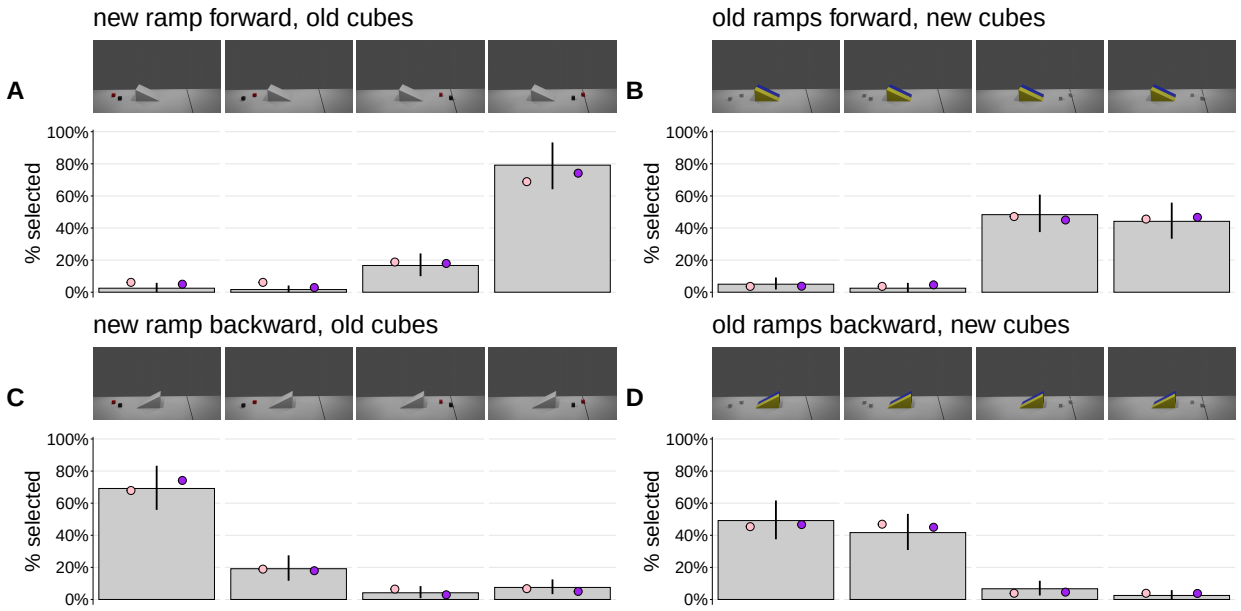

Figure 14

Experiment 3*. Participants' selections in the 'generalization task' in the 'ramp forward' condition together with the predictions of the causal abstraction model (○) and the feature-based model (●). 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 feature 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. Panels* ***A*** *through* ***D*** *show participants' selections on the four generalization trials. The trials manipulate whether the ramp is new (left), or the cubes are new (right), and whether the ramp faces forward (top), or backward (bottom).*

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

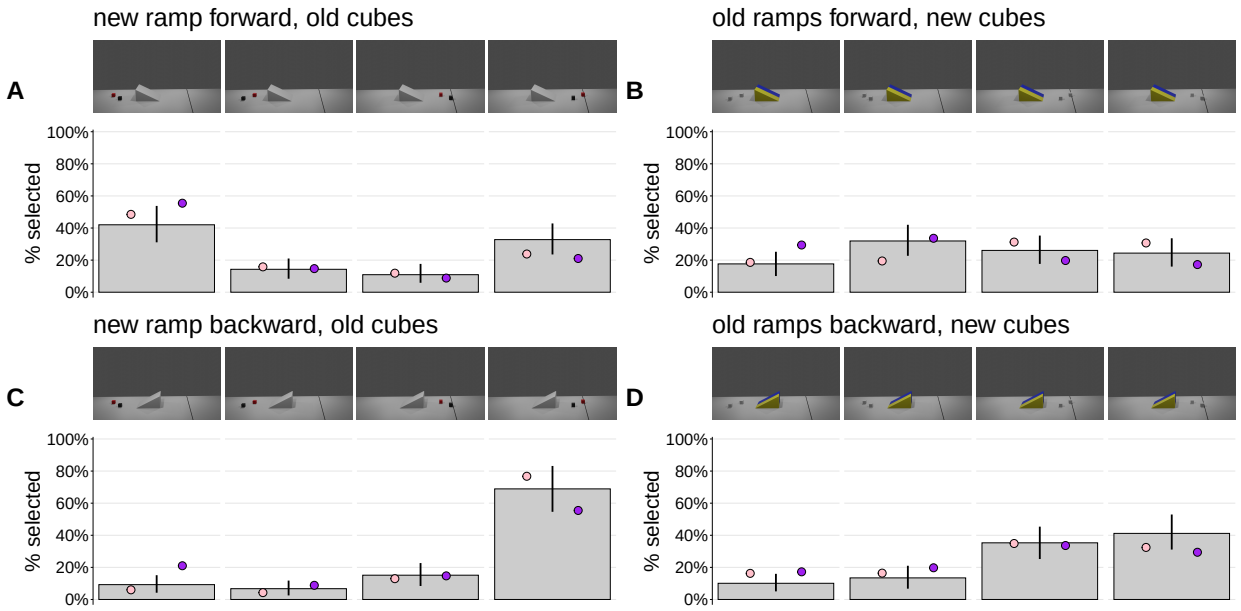

Figure 15

Experiment 3. *Participants' selections in the 'generalization task' in the 'ramp backward' condition together with the predictions of the causal abstraction model (○) and the feature-based model (●). 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 feature 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. Panels **A** through **D** show participants' selections on the four generalization trials. The trials manipulate whether the ramp is new (left), or the cubes are new (right), and whether the ramp faces forward (top), or backward (bottom).*

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 about the ramp friction than about the 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 farther 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' selections for the 'ramp backward' 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 farther 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.

The 'ramp backward' 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 backward-facing ramp, which is consistent with what they had experienced so far. In Figure 15c, most participants believed that the red cube would go farther, whereas in Figure 15d, a similar number of participants selected that one or the other cube would go farther. Again, participants were 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 farther 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 farther.

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 backward, but they make different predictions about what happens if the ramp faces forward. 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' explanations are similarly high, 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 cube would go farther. In contrast, in Figure 15b, the model is uncertain about both the side where cubes would end up, and which one would go farther.

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 $r = .97$, RMSE $= 4.89\%$ (the model predicts 32 different data

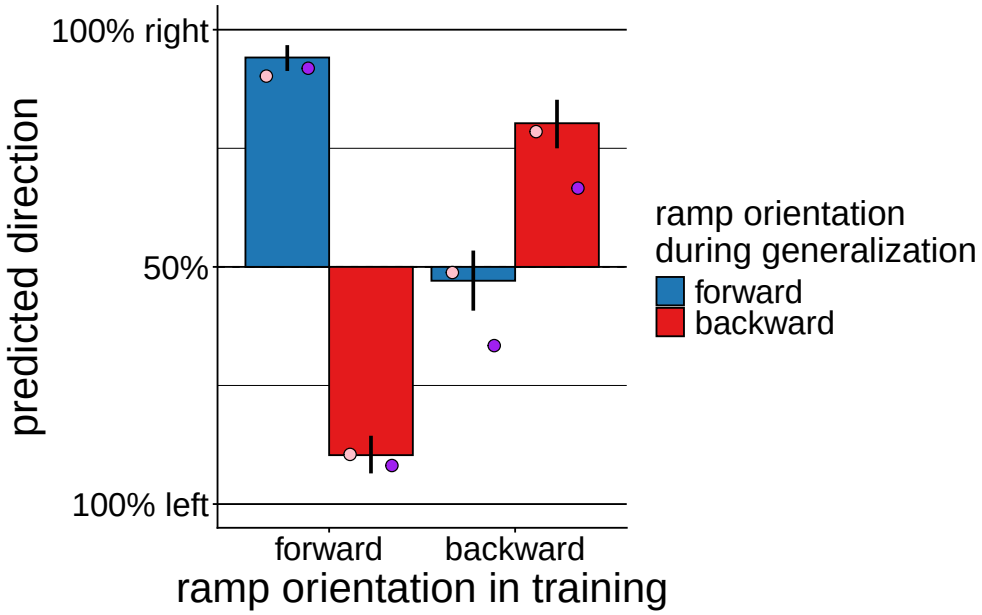

Figure 16

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.*

points shown in Figure 14 and Figure 15 using 7 free parameters; r and RMSE were computed by comparing the actual percentages with which participants selected the different response options against the predicted percentages). While the feature-based model achieves a similarly good fit to participants' selections with $r = .96$, RMSE = 6.19%, it needs 16 parameters to do so compared to the 7 parameters of the causal abstraction model. More importantly, the feature-based model fails to capture that participants' selections in the 'ramp backward' condition are asymmetric in the 'generalization task'.

Asymmetric generalization. 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, 'generalization backward' = −1). As predicted, we found a credible effect of generalization direction, $\beta = 0.90,\ 95\%\ \text{CrI} = [0.69, 1.12]$, and an interaction effect, $\beta = 1.70,\ 95\%\ \text{CrI} = [1.47, 1.95]$. While the effect of training was in the predicted direction, it wasn't credible, $\beta = -0.17,\ 95\%\ \text{CrI} = [-0.39, 0.04]$. 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.

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

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 any of the models ('other').*

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

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 any of the three explanations in the 'ramp backward' condition compared to the 'ramp forward' condition.

Explanation task

In the 'explanation task', we showed participants which of the four options they had selected for each of the 'generalization task' trials and asked them to explain, in an open-ended text box, why they had chosen that option. This task was primarily exploratory and we didn't pre-register any hypotheses as to what people would say.

We used keywords to code whether participants referred to the cubes or ramps in their responses. For example, if a response contained the words 'black', 'red', or 'white/gray cube/block/square', we coded it as a 'cube' response. If it contained the words 'blue', 'yellow/green', or 'white/gray ramp/triangle/wedge/slide/platform/slope/base', we coded it as a 'ramp' response.

Table 3 shows what proportion of participant responses talked about the cubes and the ramps. We see that participants were more likely to talk about the feature that mattered to them. So, participants for whom the cube color was diagnostic of whether it crossed the finish line were more likely to refer to the cubes rather than ramps when explaining why they chose a particular response option in the 'generalization task'. Participants for whom the ramp color mattered were slightly more likely to talk about ramps rather than cubes. Notice that, overall, participants were somewhat more likely to only talk about cubes (29%) than ramps (21%).

In addition to whether participants talked about cubes or ramps, we also coded whether they used a particular strategy. We found that 23% of responses referred to a causal mechanism, meaning that the response contained words such as 'momentum', 'gravity', 'friction', 'force', 'speed', 'launch', etc. Also, 23% of responses indicated that participants had guessed, containing words such as 'guess', 'hunch', 'not sure', etc.

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

Table 3

Experiment 3*. Proportion of participant responses in the 'explanation task' that mention both features, the cube feature, the ramp feature, or neither feature when explaining why they chose a particular answer in the 'generalization task'. Responses are separated by whether the cube feature (top) or the ramp feature (bottom) mattered for these participants in the 'prediction task'.*

	both	cube feature	ramp feature	neither
cube matters	2%	35%	16%	47%
ramp matters	3%	23%	27%	46%

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. As in Experiment 2, participants weren't more accurate on the first 'surprise task' trial when that trial matched what they had just seen as the last trial in the 'prediction task' (see Table A1).

So far, these results are consistent with the idea that people learn causal abstractions but are also consistent with a feature-based model where a linear mapping between initial and final states is 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 'generalization task' provides a critical test where the predictions of the causal abstraction model and the feature-based model come apart. Importantly, 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, participants' 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 their responses.

In the 'explanation task', where participants provided free-form responses about why they selected a particular response in the 'generalization task', we found that participants were more likely to refer to the feature that had been critical for them in the 'prediction task'. For example, participants for whom the cube color was critical were more likely to refer to cube features than to ramp features. We also found that many responses contained explicit reference to a causal mechanism, even though in this experiment, participants only saw still images of the initial and final scene, but not what happened in between.

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 go beyond simply associating predictive features with outcomes – they embody a causal story about how the observed data was generated. Crucially, the systematic nature of this inferred causal story not only guides present behavior, but also determines how individuals generalize what they have learned to novel situations in the future, in a way that a feature-based model cannot explain as well.

In Experiment 1, participants learned which objects turned on a blicket detector and which ones didn't. When only one feature mattered, such as the shape or color, participants learned this quickly. When we then surprised them and asked them what scene they had just seen, they were more likely to remember the diagnostic feature (say, the object's shape) than the non-diagnostic one (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 task'.

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 colors. 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. This pattern of results is consistent with the idea that participants learned to pay attention to the feature that was relevant for their task, and ignored the irrelevant one. This reflects a similar process of goal-dependent abstraction as in Experiment 1, albeit in a more dynamic, physical environment.

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 on 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.

To capture participants' responses in the 'generalization task' we contrasted the predictions of a causal abstraction model with those of a feature-based model. While the causal abstraction model predicts that people generalize differently depending on whether

they experienced the ramp facing forward or backward in the 'prediction task', the feature-based model predicts a symmetric pattern. Participants' responses were better captured by the causal abstraction model than the 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. Participants also often referred to aspects of the underlying causal mechanism, such as the role of friction or force, when asked to explain why they had selected a particular response in the generalization task.

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.

In the remainder, we will elaborate on three aspects that this work highlights: 1) people's propensity to create causal stories of what happened, 2) their ability to abstract the knowledge they need, and 3) the ways in which we as researchers can tell that 1) and 2) happened.

Creating stories of what happened

The idea that people create stories of what happened that go beyond what they saw or heard has a long tradition in psychology. People forget irrelevant information and systematically misremember things to fit their prior knowledge (Alba & Hasher, 1983; Bartlett, 1932; Graesser, Robertson, & Anderson, 1981). And if things don't fit, people often create ad hoc hypotheses that make them fit (S. J. Gershman, 2018; Schulz et al.,

2008). People also imbue the observable with deeper causal meaning, such as when watching geometric shapes move leads to inferences about their intentions, beliefs, and character (Heider & Simmel, 1944). One of the many places where story-telling takes center stage is the courtroom. Here, the prosecution and defense both try to tell a convincing story of what happened – one that leads the jurors to convict or acquit. Pennington and Hastie (1986) developed and tested (Pennington & Hastie, 1992) the story model for juror decision making, showing that the ease with which jurors could create a coherent story of what happened affected the judged credibility of the evidence as well as the ultimate verdict (see also Einhorn & Hogarth, 1986; Lagnado, Fenton, & Neil, 2013; Lagnado & Gerstenberg, 2017).

We also saw that many participants in Experiment 3 came up with causal stories for what happened to connect the initial and final image of the physical scene. Participants' explanations for why they thought something was going to happen in the 'generalization task' often referred to relevant aspects of the causal mechanism. The causal abstraction model is limited in that it postulates a small set of candidate hypotheses about these underlying causal mechanisms, and that it didn't come up with these explanations – we, the researchers, came up with them. More work is needed to better understand how people come up with plausible hypotheses in the first place (Phillips, Morris, & Cushman, 2019). For example, according to one proposal (Wong et al., 2025), people use their general world knowledge to determine what factors are relevant in a particular situation, and then construct a specific causal model of the situation based on their goals. Concretely, this proposal is implemented by using a large language model that translates a textual description of a task into a symbolic model (e.g., a probabilistic program; Goodman, Tenenbaum, & Gerstenberg, 2015) that is then used for reasoning. This approach combines the flexibility of large language models with the coherence of symbolic programs.

Causal abstraction through analogy and explanation

Once people have constructed the outlines of a causal story, they need to fill in the details. People's goal to explain what happened and why is likely to play a critical role here. When filling in causal stories, we have to be selective. Sometimes, we can use relevant causal knowledge from a related domain to figure out what matters. Such reasoning by analogy has been demonstrated in decades of work (e.g., Gentner, 1983; Gentner & Asmuth, 2019; Gentner & Markman, 1994). By mapping the abstract structure from one situation onto another, a reasoner can often quickly find solutions to novel problems.

In addition to analogical reasoning, attempting to explain why something happened is likely to facilitate causal abstraction, too (Lombrozo & Liquin, 2023; Walker, Lombrozo, Legare, & Gopnik, 2014). The feeling of failing to give a satisfactory explanation can motivate subsequent search, something that Liquin and Lombrozo (2020) call explanation-seeking curiosity. Importantly, this need not necessarily happen in a communicative context, where one person asks another to explain. It can happen within the same person, too (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Lombrozo, 2024).

In Experiment 3, we saw evidence that participants inferred different causal stories of what happened when the cube ended up on the other side of a backward-facing ramp. What story participants created subsequently affected how they generalized their knowledge to a novel situation. In related work, Legare (2012) asked children aged 2 to 6 years why something happened. When presented with inconsistent evidence, a child's explanation systematically affected how they interacted with the scene and generated new hypotheses (see also Ahn & Bailenson, 1996; Walker, Lombrozo, Williams, Rafferty, & Gopnik, 2017). In Experiment 3, we only asked participants to explain why they had selected a particular response in the 'generalization task' at the very end of the experiment. It would be interesting to see what would happen if participants were asked to generate explanations earlier on. One possibility, consistent with prior literature, is that this would lead to even faster causal abstraction.

Probing mental representations

Causal abstraction happens in the mind. However, as researchers, we only have access to brain signals and behavioral signatures. Both provide important insights. For example, brain signals can help shed light on whether the process of abstraction occurs during encoding (Shohamy & Wagner, 2008) or retrieval (Favila, Lee, & Kuhl, 2020), or both (Schlichting, Mumford, & Preston, 2015), and can help localize where in the brain this information is processed (Bernardi et al., 2020).

We focused here on behavioral signatures of abstraction. Specifically, we assessed how participants' ability to predict what will happen changed over time, what they remembered, and how they generalized what they learned to new situations. All three measures provide evidence for causal abstraction. For example, while participants' predictions in Experiment 1 became more accurate when they had more opportunity to learn, their memory of what just happened became more inaccurate (compared to a group of participants who had received less training). And, in Experiment 3, participants made systematic mistakes in the 'surprise task' that then translated into how they generalized their knowledge to a novel situation. Additional behavioral signatures, such as reaction times, eye movements, and confidence measures (e.g., Ashby & Maddox, 2005; Gerstenberg et al., 2017; Ho et al., 2022; Nissen & Bullemer, 1987), may provide further insight into the underlying cognitive mechanisms that support causal abstraction. For example, by tracking participants' eye movements in the 'ramp task', one could provide more direct evidence for selective attention to the cubes versus the ramps depending on which aspect of the physical scene mattered more for the 'prediction task'. Reaction time data could reveal the benefits and costs of abstraction: we would expect fast reaction times when the learned abstraction fits the task, and slow reaction times when there is a mismatch.

Limitations and future directions

We studied causal abstraction in a small set of situations involving blicket detectors and cubes sliding down ramps. Future work should study causal abstraction in more

complex environments. This is challenging from a methodological perspective because there could be many different causal stories that come to people's minds. Causal Bayesian networks and structural equation models have proven useful as flexible tools for representing abstract causal knowledge (Pearl, 2000; Sloman, 2005; Spirtes, Glymour, & Scheines, 2000). However, these tools aren't particularly well-suited for representing spatiotemporal information, which is critical for modeling physical systems (Jensen, 2019). And while physics engines are great for representing spatiotemporal information, it's unclear how they can be used to represent more abstract knowledge (Ullman et al., 2017). More work is needed to bridge that gap: models that can flexibly represent the world at different levels of abstraction (Wong et al., 2025).

In Experiment 3, the causal abstraction model stipulated three 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. One way of doing so would be to directly link the noise distribution in the physical simulation model with the learning input. The model would start with a wide noise distribution over the physical parameters and reduce that distribution with each new data point, weighted by how much that parameter seems to matter for the task at hand.

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. There are limits to what behavioral tests can reveal about people's mental models. Future work may explore neuroscientific evidence, too. This may help, for example, with teasing apart whether the compression of causally irrelevant features happens primarily during the encoding stage or the retrieval stage of information processing.

Finally, the way in which we probed people's knowledge in our tasks highlights the costs of causal abstraction. People misremember things that they had seen earlier, and they are uncertain about what would happen in novel situations where a feature they had earlier learned to ignore now matters. More work is needed that illustrates the benefits of learning causal abstractions. These could include fast decision times for complex scenes, and robust generalization across causally similar environments.

Constraints on generality

Our conclusions are based on samples of adults recruited online through Prolific, restricted to fluent English speakers residing in the United States with high approval ratings on the platform. Across the three experiments, the samples were predominantly White and included somewhat more women than men (see the 'Participants' sections for each experiment). We have no specific reason to expect that the core phenomenon we studied – that people construct goal-dependent causal abstractions that shape memory, prediction, and generalization – is restricted to this population, since the underlying cognitive processes (selective attention, memory reconstruction, and causal inference) are thought to be widespread across human cognition and have been documented in other populations, including young children (Legare, 2012). That said, we did not test for effects of age, culture, or domain expertise, and the paradigms we used (a blicket detector and a ramp-and-cube physics task) are simplified laboratory analogs of causal learning; the specific rate and pattern of abstraction we observed may depend on task complexity, the amount of training, and the presence of performance-based incentives, all of which could

differ in more naturalistic settings.

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

Author contributions

Steven Shin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Project administration, Writing – review & editing. Chuqi Hu: Methodology, Software, Validation, Data curation, Writing – review & editing. Paul S. Muhle-Karbe: Conceptualization, Supervision, Writing – review & editing. Tobias Gerstenberg: Conceptualization, Methodology, Software, Validation, Formal analysis, Visualization, Supervision, Project administration, Funding acquisition, Writing – original draft, Writing – review & editing.

Appendix A

Order effects in Experiments 2 and 3

We explored whether participants were more accurate at the first trial in the 'surprise task' when that type of trial matched the last trial they had seen on the 'prediction task'. Remember that the trial order was randomized both for the 'prediction task' and the 'surprise task'. This means that it's possible for a participant to just have seen a red cube and a blue ramp as the last prediction trial, for example, and then see that exact same type of trial as the first one in the 'surprise task'.

We found that neither in Experiment 2 nor in Experiment 3 were participants more accurate when the first 'surprise task' trial matched the last 'prediction task' trial compared to when it mismatched (see Table A1). The log-odds ratio for the contrast 'different – same' across all conditions in Experiment 2 was $\beta = -0.250,\ 95\%\ \text{CrI} = [-0.752, 0.209]$, and $\beta = 0.185,\ 95\%\ \text{CrI} = [-0.403, 0.807]$ in Experiment 3.

Table A1

Experiments 2 and 3*. Accuracy on the first 'surprise task' trial based on whether the last 'prediction task' trial was 'different' or the 'same'. Overall, whether the last trial was 'different' or the 'same' had no credible effect on accuracy.*

experiment	condition	trial	accuracy	n
2	long	different	0.54	90
2	long	same	0.63	30
2	short	different	0.45	85
2	short	same	0.57	35
2	conjunctive	different	0.59	85
2	conjunctive	same	0.56	34
3		different	0.44	183
3		same	0.39	56

Appendix B

Detailed description of the causal abstraction model

Here, we provide more details on how some of the computations in the model are performed. The causal abstraction model considers three different hypotheses of how the data may have been generated (see Figure 11). Two free parameters, p and q, characterize the prior distribution over these hypotheses.

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

$$p(\text{hypothesis}_{\text{physics}}|\text{cube position}) = \frac{p(\text{cube position}|\text{hypothesis}_{\text{physics}}) \cdot p(\text{hypothesis}_{\text{physics}})}{\sum_{i\in\{\text{physics,agent,ramp}\}} p(\text{cube position}|\text{hypothesis}_i) \cdot p(\text{hypothesis}_i)}, \tag{B1}$$

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.

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 be fully ruled out even if a cube ended up on the right of a backward-facing ramp. As illustrated in Figure 11b, if the ramp faces forward, the 'physics' and 'agent' hypotheses 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' hypotheses 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

$$p(\text{option}_i|\text{data}) = \sum_{j \in \{\text{physics,agent,ramp}\}} p(\text{option}_i|\text{hypothesis}_j) \cdot p(\text{hypothesis}_j|\text{data}), \tag{B2}$$

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 $p(\text{hypothesis}_j|\text{data})$ is the posterior probability of that hypothesis based on the data seen so far (computed in Equation B1).

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. We made sure that an object's friction was positive after the noise was added. If a high noise value resulted in a negative friction, we resampled that noise value. 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. If a cube ended up on the wrong side of the ramp due to a high noise value, we resampled that noise value. For example, if a cube's trajectory was simulated according to the agent hypothesis on a backward-facing ramp (see Figure 11b) but ended up on the left side (because of a high ramp friction value), we reran that simulation.

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 B2 into a

discrete choice, the model uses a softmax function

$$p(\text{option}_i) = \frac{\exp(\beta \cdot p(\text{option}_i|\text{data}))}{\sum_{j=1}^{4} \exp(\beta \cdot p(\text{option}_j|\text{data}))}, \tag{B3}$$

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 7 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: $p = 0.96$, $q = 0.37$, $\epsilon = 0.015$, $\sigma_{\text{cube}} = 0.09$, $\sigma_{\text{ramp}} = 1.00$, $\kappa = 1.00$, and $\beta = 600,257$.[8]

[8]The beta parameter is large because the softmax in Equation B3 uses values from the posterior predictive distribution as inputs (see Equation 2) and these values are very small.

References

Ahn, W.-k., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, *31*(1), 82–123.

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*(2), 203.

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive science*, *26*(1), 39–83.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu. Rev. Psychol.*, *56*(1), 149–178.

Balaban, H., & Ullman, T. D. (2025). The capacity limits of moving objects in the imagination. *Nature Communications*, *16*(1), 5899.

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

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge university press.

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).

Belledonne, M., Butkus, E., Scholl, B. J., & Yildirim, I. (2026). Adaptive computation as a new mechanism of dynamic human attention. *Psychological Review*, *133*(3), 534.

Beller, A., & Gerstenberg, T. (2026). Causation, meaning, and communication. *Psychological Review*, *133*(2), 339–381.

Beller, A., Xu, Y., Siegel, M., Grant, S., Brown, A., Tenenbaum, J. B., ... Gerstenberg, T. (2025). Multimodal inference through mental simulation. *PsyArXiv*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, *183*(4), 954–967.

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., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301.

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880–1910.

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.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576.

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

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.

Chen, T., Cheyette, S., Allen, K., Tenenbaum, J., & Smith, K. (2026). "just in time" world modeling supports human planning and reasoning. *arXiv preprint arXiv:2601.14514*.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–382.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182.

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, *120*(1), 190.

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.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3–19.

Favila, S. E., Lee, H., & Kuhl, B. A. (2020). Transforming the concept of memory reactivation. *Trends in neurosciences*, *43*(12), 939–950.

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.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., ... Icard, T. (2025). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, *26*(83), 1–64.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy*. *Cognitive science*, *7*(2), 155–170.

Gentner, D., & Asmuth, J. (2019). Metaphoric extension, relational categories, and abstraction. *Language, Cognition and Neuroscience*, *34*(10), 1298–1307.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological science*, *5*(3), 152–158.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American psychologist*, *52*(1), 45.

Gershman, S., Cohen, J., & Niv, Y. (2010). Learning to selectively attend. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 32).

Gershman, S. J. (2018, May). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50.

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., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744.

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.

Graesser, A. C., Robertson, S. P., & Anderson, P. A. (1981). Incorporating inferences in narrative representations: A study of how and why. *Cognitive Psychology*, *13*(1), 1–26.

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

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

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

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.

Holland, P. C., & Schiffino, F. L. (2016). Mini-review: Prediction errors, attention and associative learning. *Neurobiology of learning and memory*, *131*, 207–215.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*, 135–163.

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.

Icard III, T. F., & Goodman, N. D. (2015). A resource-rational approach to the causal frame problem. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 37).

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

Jensen, D. (2019). Overcoming the poverty of mechanism in causal models. In S. Kleinberg (Ed.), *Time and causality across the sciences.* Cambridge 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.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*(1), 227–254.

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.

Konidaris, G. (2019). On the necessity of abstraction. *Current opinion in behavioral sciences*, *29*, 1–7.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–248.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning.

Journal of mathematical psychology, *45*(6), 812–863.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759.

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

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation*, *4*(1), 46–63.

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451–460.

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.

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child development*, *83*(1), 173–185.

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.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological bulletin*, *142*(10), 1111.

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.

Liquin, E. G., & Lombrozo, T. (2020, Jun). A functional approach to explanation-seeking curiosity. *Cognitive Psychology*, *119*, 101276.

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

Lombrozo, T. (2024). Learning by thinking in natural and artificial minds. *Trends in Cognitive Sciences*.

Lombrozo, T., & Liquin, E. G. (2023). Explanation is effective because it is selective. *Current Directions in Psychological Science*, *32*(3), 212–219.

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. *Oxford handbook of causal reasoning*, 415–432.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological review*, *111*(2), 309.

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.

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature communications*, *11*(1), 46.

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological review*, *82*(4), 276.

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

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

Mienye, I. D., & Swart, T. G. (2025). Deep autoencoder neural networks: A comprehensive

review and new perspectives: Id mienye, tg swart. *Archives of computational methods in engineering*, *32*(7), 3981–4000.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

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. *Neuron*, *111*(23), 3885–3899.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive psychology*, *19*(1), 1–32.

Niv, Y. (2019). Learning task-state representations. *Nature neuroscience*, *22*(10), 1544–1553.

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.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

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

Pearce, J. M., & Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532.

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

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, *51*(2), 242.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for

juror decision making. *Journal of Personality and Social Psychology*, *62*(2), 189–206.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*, *23*(12), 1026–1040.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

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

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature communications*, *6*(1), 8151.

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, *91*(6), 1402–1412.

Schuck, N. W., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P. A., Haynes, J.-D., & Reverberi, C. (2015). Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron*, *86*(1), 331–340.

Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, *364*(6447), eaaw5181.

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., & Jenkins, A. C. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*(2), 211–223.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In *The psychology of learning and motivation* (Vol. 21, pp. 229–261). San Diego, CA: Academic Press.

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

Sharpe, M. J., Stalnaker, T., Schuck, N. W., Killcross, S., Schoenbaum, G., & Niv, Y. (2019). An integrated model of action selection: distinct modes of cortical control of

striatal decision making. *Annual review of psychology*, *70*(1), 53–76.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron*, *60*(2), 378–389.

Simon, H. A. (1979). *Models of thought* (Vol. 352). Yale university press.

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.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Smith, K. A., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and retrodiction. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1467–1472). Austin, TX: Cognitive Science Society.

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.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. The MIT Press.

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).

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge University Press.

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.

Tian, L. Y., Garzón Gupta, K., Hanuska, D. J., Rouse, A. G., Eldridge, M. A., Schieber, M. H., ... Freiwald, W. A. (2026). Neural representation of action symbols in primate frontal cortex. *Nature*, 1–11.

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.

Unger, L., & Sloutsky, V. M. (2023). Category learning is shaped by the multifaceted development of selective attention. *Journal of Experimental Child Psychology*, *226*, 105549.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, *82*(1), 27–58.

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.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General*, *124*(2), 181.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts

children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child development*, *88*(1), 229–246.

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.*

Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in human neuroscience*, *5*, 189.

Wong, L., Collins, K. M., Ying, L., Zhang, C. E., Weller, A., Gerstenberg, T., ... Brooke-Wilson, T. (2025). Modeling open-world cognition as on-demand synthesis of probabilistic models. *arXiv*.

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.

Xia, L., & Collins, A. G. (2021). Temporal and state abstractions for efficient learning, transfer, and composition in humans. *Psychological review*, *128*(4), 643.

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

Yildirim, I., Gerstenberg, T., Saeed, B., Toussant, M., & Tenenbaum, J. B. (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic

constraints. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3584–3589). Austin, TX: Cognitive Science Society.

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.

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023). Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*, *152*(8), 2237–2269.