Learning what matters: Causal abstraction in human inference

Steven Shin (Steven.M.Shin.23@dartmouth.edu)
Department of Cognitive Science, Dartmouth University

Tobias Gerstenberg (gerstenberg@stanford.edu)
Department of Psychology, Stanford University

Abstract

What shape do people's mental models take? We hypothesize that people build causal models that are suited to the task at hand. These models abstract away information to represent what matters. To test this idea empirically, we presented participants with causal learning paradigms where some features were outcome-relevant and others weren't. In Experiment 1, participants had to learn what objects of different shape and color made a machine turn on. In Experiment 2, they had to predict whether blocks sliding down ramps would cross a finish line. In both experiments, participants made systematic errors in a surprise test that asked them to recall what they had seen earlier. The errors people made suggest that they had built mental models of the task that privileged causally relevant information. Our results contribute to recent efforts trying to characterize the important role that causal abstraction plays in human learning and inference.

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

Introduction

When modeling the world, we can't account for everything. So how do we choose what to represent? Here is a simple idea: we represent what we need. When building mental models of the world, our representations are tailored to the task at hand, abstracting away much of the information that's available in principle. Recent work suggests that people's mental models of the physical world may in important respects be similar to the kinds of physics engines that are used in modern day computer games (Kubricht, Holyoak, & Lu, 2017; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). These physics-engine-fueled models quantitatively capture people's predictions about the future (Battaglia, Hamrick, & Tenenbaum, 2013; Smith & Vul, 2012), inferences about the past (Beller, Xu, Linderman, & Gerstenberg, 2022; Smith & Vul, 2014), and causal judgments about what happened (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Gerstenberg & Tenenbaum, 2017; Zhou, Smith, Tenenbaum, & Gerstenberg, 2022). While encouraging, there is also a sense in which these models aren't quite right (Ludwin-Peery, Bramley, Davis, & Gureckis, 2020). For a physics engine to simulate an object it needs to know its exact location, mass, shape, friction, etc. However, people's mental models may not represent the world at this level of detail (Davis & Marcus, 2016), and they may only mentally simulate some aspects but not others (Bass, Smith, Bonawitz, & Ullman, 2022).

Abstract causal models

Mental models can be formulated at multiple levels of abstraction (Griffiths & Tenenbaum, 2009). While we ultimately have to take actions in continuous space and time, we often don't think about the world that way. Instead, our mental models abstract away many of the lower level details (e.g. Beckers, Eberhardt, & Halpern, 2020; Beckers & Halpern, 2019; Beller, Bennett, & Gerstenberg, 2020; Chalupka, Eberhardt, & Perona, 2017; Gerstenberg et al., 2021). How do we choose the right level of abstraction (Gerstenberg & Stephan, 2021; Halpern & Hitchcock, 2011; Woodward, 2015)?

One proposal from philosophy suggests that the variables in a model should be 'proportional' to one another (Woodward, 2021; Yablo, 1997): cause and effect variables should be specified at a level of detail that matches. For example, when representing the relationship between an ordinary light switch and a light bulb, binary variables will do just fine. However, to capture the relationship between a dimmer and a dimmable light bulb, we would want continuous variables instead. What we wouldn't want would be a continuous variable for a switch capturing its exact position over time, when the bulb is only ever on or off. 'Proportionality' provides a useful constraint on specifying the values of variables in a causal model. However, it doesn't yet answer the question of what variables to include. Here, we need to consider the agent's goals (Wellen & Danks, 2014).

Goal-dependent mental representations

People's goals constrain what visual information they attend to (Maruff, Danckert, Camplin, & Currie, 1999). Because our cognitive resources are limited, we need to allocate them efficiently (Bates, Lerch, Sims, & Jacobs, 2019; Brady & Tenenbaum, 2013). For example, when people plan how to navigate a maze, they build simplified representations that only contain what matters at a fine level of detail (Ho et al., 2022). When asked to recall where an obstacle was located, they remember it well when the obstacle affected their planned route, but less so when the obstacle didn't matter for their plan. As Ho (2019) argue, there is value in abstraction. Good abstractions help us to learn and transfer that knowledge to new tasks.

Goal-dependent causal models

Prior work on abstraction in causal models has been mostly theoretical (Beckers & Halpern, 2019; Chalupka et al., 2017).

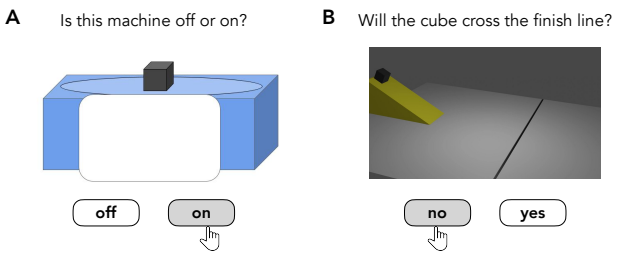


Figure 1: Example trial of the 'prediction task' in the blicket experiment (**A**) and physics experiment (**B**).

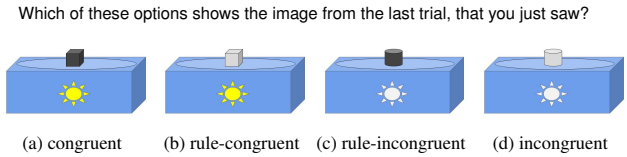


Figure 2: **Experiment 1**. In this example of the 'surprise test' the cubes are blickets and the cylinders aren't. For the labels, we assume that the black cube was shown last in the 'prediction task' (see Figure 1a).

And prior work on goal-dependent representations has focused on visual attention (Maruff et al., 1999), or planning in navigation (Ho et al., 2022). Here, we bring these two strands of research together. We ask whether people build goal-dependent causal models that are suited to the task at hand. In Experiment 1, participants perform a simple causal learning task and, when asked in a surprise test what they just saw, make systematic memory errors favoring causally relevant information (see Figure 1a). Experiment 2 features a physical prediction task. Again, we find systematic errors in a surprise test that are consistent with a goal-dependent causal model of the task (see Figure 1b). 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 of the pre-registrations, data, and materials here: `https://github.com/cicl-stanford/abstract_causation`

Experiment 1: Causal abstraction of blickets

In this experiment, we use the popular blicket detector paradigm (see, e.g. Gopnik et al., 2004; Sobel & Kirkham, 2006) to investigate whether people build task-dependent causal abstractions.

Methods

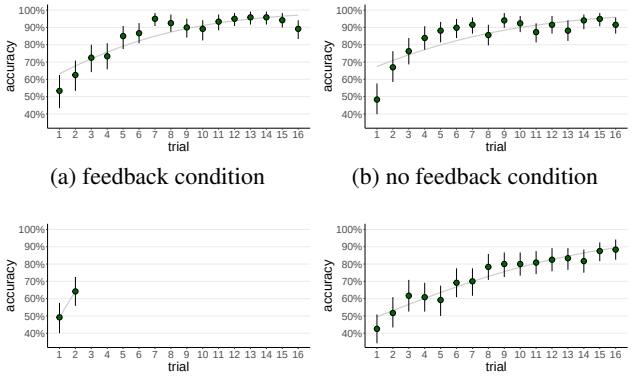
Participants A total of 482 participants were recruited through Prolific (*age*: M = 38, SD = 14; *gender*: 267 female, 192 male, 12 non-binary, 11 other or no response; *race*: 368 White, 45 Asian, 30 Black, 27 Multiracial, 2 Hispanic, 1 Native, 1 White African, and 8 no response) took part in the four experimental conditions (*feedback*: $N = 124$, *no feedback*: $N = 118$, *short*: $N = 120$, *conjunctive*: $N = 120$). All participants were based in the US, fluent in English, and had approval ratings of at least 95% with 10 or more prior submissions. A target sample size of 120 participants was selected for each of the four conditions, based upon a frequentist power analysis using a significance threshold of $p = .05$ and a target power of 80%. Participants were compensated with both a base payment, and a performance bonus based upon their overall accuracy within the task. The average compensation exceeded $14/hr in each condition.

Procedure The stimuli in this experiment showed a 'blicket machine' with one of four objects placed on top of the machine: a dark cube, a light cube, a dark cylinder, or a light cylinder. Figure 2 shows an example of the blicket detector in action. The machine would 'turn on' (as indicated by the sun turning yellow) whenever blickets were placed atop the machine. The diagnostic feature of a 'blicket' was either its shape or color, and was counterbalanced between participants. In Figure 2, shape is the diagnostic feature, while color is irrelevant. Here, cubes (of any color) are blickets while cylinders (of any color) are not. For other participants, color was diagnostic of blickets, while shape was irrelevant.

At the beginning of the experiment, participants were familiarized with the 'blicket machine' and informed that some but not all objects would make the machine turn on. Participants were then given a brief comprehension check to ensure that they understood how the machine 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 machine, with one of the four objects placed atop the machine like shown in Figure 1a. The status of the machine (whether it was 'on' or 'off') was hidden by a white occluder. Participants were asked to indicate whether the machine was 'on' or 'off'. After responding, the occluder was removed, the status of the machine revealed, and participants received feedback about whether their response was correct. Participants received a bonus for each correct response on a 'prediction trial'.

After the last 'prediction trial', participants viewed a 'surprise test' asking them to recall which of the four objects they had seen in the preceding trial. Participants clicked on the image in a 2×2 grid that they believed they had seen last.

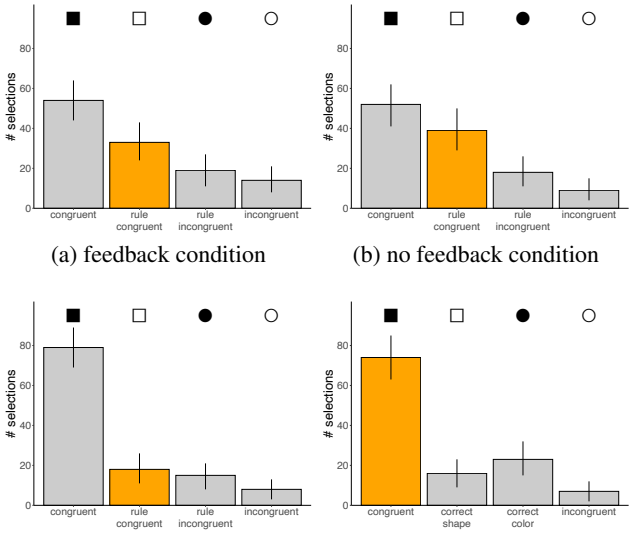
Design The experiment had four conditions. In the 'feedback' condition, participants received feedback on the final prediction trial, indicating whether they responded correctly, before being presented with the surprise test. In the 'no feedback' condition, participants didn't receive feedback on the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



(a) feedback condition (b) no feedback condition

(c) short condition (d) conjunctive condition

Figure 3: **Experiment 1**. Accuracy in the prediction task. Lines show logistic regression model fits. Error bars in all figures show 95% bootstrapped confidence intervals.



(a) feedback condition (b) no feedback condition

(c) short condition (d) conjunctive condition

Figure 4: **Experiment 1**. Participants' selections in the 'surprise test'. 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).

Predictions

Figure 2 shows the different response options and the corresponding labels we give to each option. We call responses 'congruent' when they correctly identify the object from the preceding trial (here, the black cube). Responses are 'rule-congruent' when they would have led to the same outcome as the correct response. Here, the light cube is rule-congruent because it is also a blicket. The black cylinder is 'rule-incongruent' because it shares its color with the correct response, but that feature is not causally relevant. 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 light cube), and one incongruent option.

We predicted that, as participants learned what distinguishes blickets from non-blickets, their representations of the stimuli 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 shape of the object than its color. As a consequence, if participants made a mistake in recalling what they just saw, they would be more likely to select the rule-congruent rather than the rule-incongruent option (and least likely to select the incongruent option). In the 'feedback' condition, one might worry that participants would only remember the feedback they received (e.g. that the blicket detector was on) but not the object they saw. 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 a 'no feedback' condition where participants didn't get feedback on the final prediction trial before the surprise test. If feedback was driving the effect, we would expect any difference in selections to disappear in that condition. We predicted that, in the 'short' condition, participants would not have enough evidence to learn the relevant causal structure, and would therefore not have a privileged memory of the causally relevant feature. Thus, we predicted that they would be just as likely to select the rule-congruent and rule-incongruent option. Similarly, in the 'conjunctive' condition, because both the color and the shape of the objects are causally relevant, participants would not have a privileged memory of either feature over the other, and thus would be equally likely to select each partially congruent option.

Results

Prediction task Figure 3 shows participants' accuracy in the 'prediction task' across trials. Participants were able to quickly learn which objects were blickets. In the conditions in which one feature mattered, more than 80% of participants successfully predicted whether or not the blicket detector was 'on' by trial 5. In the conjunctive condition (Figure 3d), it took participants a little longer to learn the rule.

Surprise test Figure 4 shows participants' selections in the 'surprise test' for each condition. Results are aggregated over the counterbalanced conditions. For the purpose of visualization, we assume that cubes are blickets and that the black cube was shown immediately prior to the surprise test. Our main hypothesis was that if people misremembered what they had just seen, they'd be more likely to select the rule-congruent option than the rule-incongruent option. In the 'feedback' condition in which participants saw the outcome on their final prediction trial, participants were more likely to select the rule-congruent than the rule-incongruent option but the difference was not significant, $B = 0.55$, 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 $B = 0.77$, 95% CI $[0.21, 1.33]$, $p = .01$.

In the 'short condition' we correctly predicted that there would be no significant difference in selecting the rule-congruent versus rule-incongruent option, $B = 0.18$, 95% CI $[-0.5, 0.87]$, $p = .6$. Interestingly, participants were more likely to respond correctly here compared to the longer conditions, even though they had seen many fewer prediction trials. Finally, in the 'conjunctive' condition, we correctly predicted that there would be no difference in selecting either of the two partially matching responses, $B = -0.36$, 95% CI $[-1, 0.28]$, $p = .26$.

Discussion

Participants had no trouble learning what distinguished blickets from non-blickets. However, learning this simple rule had a consequence: participants didn't encode all the information about the stimulus, but paid specific attention to those features that were causally relevant. This was revealed through the systematic errors they made when asked to recall what they had just seen. For example, when shape (but not color) was diagnostic of blickets, participants' incorrect responses were more likely to have the same shape as the correct response, than they were to have the same color.

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. This is addressed in Experiment 2.

When participants weren't given time to build a simplified causal representation in the 'short' condition, they were more likely to correctly select the object they had last seen, and when they made a recall error, they were just as likely to select rule-congruent and rule-incongruent objects. Similarly, in the 'conjunctive' condition, when both shape and color mattered, participants were again more accurate in recalling the object they had last seen, and were equally likely to select either of the partially congruent options.

Experiment 2: Causal abstraction of physics

Experiment 1 provides evidence for the role of abstraction in causal learning using a simple 'blicket detector' task. In Experiment 2, we extend these findings to the domain of intuitive physical reasoning (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). In this experiment, participants were asked whether a block sliding down a ramp would would cross a finish line (see Figure 1b). Like in Experiment 1, we manipulated two features: the color of the block and the color of the ramp. Each feature was diagnostic for the friction associated with the object.

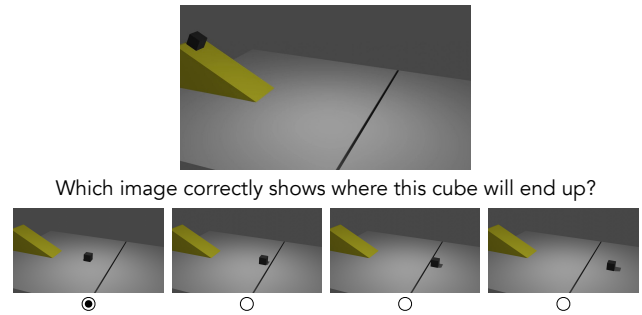


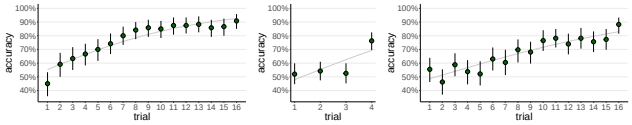
Figure 5: **Experiment 2**. In the 'surprise test', participants selected which image shows where the block will end up.

This physical setting expands the 'blicket detector' task in a number of ways. First, on a fine level of granularity, the outcome now features four possible states: the different positions of where the block ends up (see Figure 5). On a more abstract level, however, there are only two outcome states: whether or not the block crosses the finish line. This allows us to test whether people map a larger space of possible outcomes onto a smaller space that matters for their goal. Second, we can show physically realistic animations of what happens in each scenario. The setting thus comes a little closer to the kinds of situations we may experience in our everyday lives. Finally, this task addresses a potential confound from Experiment 1. This time, the predicted pattern of results cannot be explained by participants' memory of their response immediately preceding the surprise test. Instead of asking participants what they saw on the last trial, we ask participants to make predictions about where exactly the block will end up (see Figure 5). This task has the added advantage of allowing us to get more data from each participant: four judgments instead of one.

Methods

Participants Participants were recruited through Prolific using the same inclusion criteria as in Experiment 1. 359 participants (*age*: M = 38, SD = 15; *gender*: 186 female, 161 male, 9 non-binary, 3 no response; *race*: 272 White, 32 Black, 29 Asian, 18 Multiracial, 3 Native, 5 other or no response) 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/hr in each condition.

Procedure The stimuli for this experiment consisted of simple videos showing a block sliding down a ramp and then along a plane. Some blocks would slide beyond a finish line, while others would stop short. Blocks were either red or black and ramps were either blue or yellow. The color of a ramp, or of a block, was diagnostic for its surface friction. After sliding down the ramp, and along the plane, a block would stop in one of four equally spaced positions. The first and second positions did not cross the finish line, while the third and fourth positions did. Importantly, the surface frictions were set such that the friction of either the block or the ramp determined



(a) long condition (b) short (c) conjunctive condition

Figure 6: **Experiment 2**. Accuracy in the prediction task. Lines show logistic regression model fits.

whether the block would cross the finish line. For example, in the clips shown at the top of Figure 7, red blocks cross the finish line and black blocks don't. Here, blocks slide further on blue compared to yellow ramps. So, although both the ramp and the block contribute to the block's final position, attending to the block alone is sufficient for predicting whether it crosses the finish line.

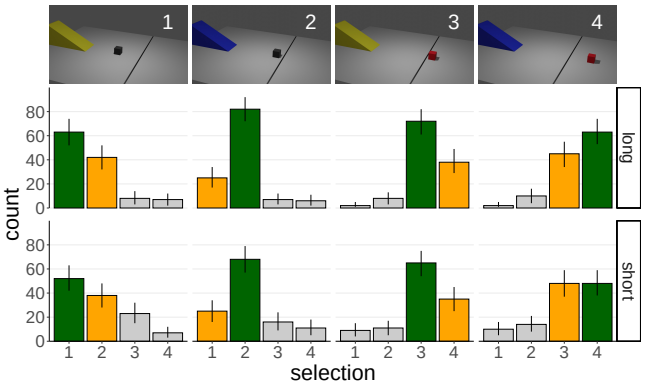
Participants were familiarized with the scene, and were instructed that they would need to make predictions about whether the block would cross the finish line in each of the four possible scenarios. Participants were then given a brief comprehension check, and were informed that they would receive a bonus for each correct prediction. The main body of the experiment then consisted of a set of 'prediction trials' in which participants were presented with an image showing a block at the top of a ramp like in Figure 1b. Participants were asked if the block would cross the finish line. Upon responding, participants were shown a video of the block sliding down the ramp and coming to rest. They received feedback about whether their response was correct.

After completing the final 'prediction trial', participants were presented with a 'surprise test'. Now, rather than indicating whether the block would cross the finish line, participants were asked: "Which image correctly shows where this cube will end up?" like in Figure 5. Participants responded by selecting one of the four images. The 'surprise test' included one trial for each of the four scenarios (black/red block on yellow/blue ramp). The order of these trials was randomized, and participants received no feedback on these trials.

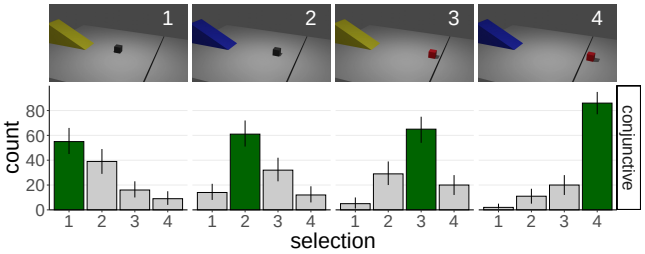
Design This experiment had three conditions. In the 'long' condition, participants completed 16 prediction trials. In the 'short' condition, participants completed only 4 prediction trials (one for each combination of block and ramp). Finally, in the 'conjunctive' condition, the finish line was moved such that it fell between the third and fourth positions. As a result, whether the block crossed the finish-line now depended on both the friction of the block, and that of the ramp. In the 'conjunctive' condition participants completed 16 prediction trials. The causally pivotal object (block or ramp) as well as the colors which corresponded to high or low frictions for each object were counterbalanced between participants.

Predictions

In the 'long' condition, we predicted that participants' selections across all four trials would be more strongly affected



(a) long condition (top) and short condition (bottom)



(b) conjunctive condition

Figure 7: **Experiment 2**. Participant selections of different end positions in the 'surprise test', separated by the correct response. Green bars indicate the correct response, orange bars show outcome-congruent responses.

by rule-relevant features compared to rule-irrelevant features. 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. While both of these positions are equidistant from position 2, they fall on different sides of the finish line and would thus lead to different outcomes.

In the 'short' condition, we predicted that there would be no difference in how strongly rule-relevant and rule-irrelevant features affected participants' selections, and that they would be just as likely to select outcome-congruent or outcome-incongruent responses when the block's correct final position was 2 or 3.

Finally, in the 'conjunctive' condition, we predicted that, when the correct position was 3, participants would be more likely to selection position 2 (which would lead to the same outcome) than position 4. We also predicted that participants would be more likely to select the correct response when the ground truth was position 4 than for the other three positions.

Results

Prediction task Figure 6 shows participants' accuracy in predicting whether the block would cross the finish line over

the course of the prediction test trials. Somewhat surprisingly, participants' accuracy in the short condition was quite high on the fourth and final trial. Like in Experiment 1, participants found it more difficult to learn the conjunctive rule.

Surprise test We will discuss the selection results in the 'surprise test' from the three conditions in turn.

Long condition. Figure 7a (top) shows participants' selections in the 'long' condition for each of the four combinations of blocks and ramps. For visualization purposes, we assume here that black blocks don't cross the finish line whereas red blocks do, and that blocks slide further on blue compared to 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 block would end up in position 2, the outcome-congruent response would be position 1 because for both positions, the block would not have crossed the finish line.

To test the prediction that 'relevant' features affect participants' selections more strongly than 'irrelevant' features, we ran a Bayesian ordinal mixed effects regression with 'relevant' and 'irrelevant' feature plus their interaction as fixed effects, and random intercepts for participants.[1] 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 'relevant' than by 'irrelevant' features, $M = 0.85$, 95% highest posterior density interval (HDI) $= [0.7, 1.01]$.

To test the prediction that participants are more likely to select outcome-congruent responses when the ground truth final 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, we found that participants were more likely to select incorrect responses that were outcome-congruent than ones that were outcome-incongruent, 90% [75%, 99%].

Short condition. Figure 7a (bottom) shows participants' selections in the 'short' condition. In contrast to what we predicted, 'relevant' features again had a stronger influence on participants' selections than 'irrelevant' features, 0.53 [0.39, 0.67], though this difference was smaller than in the 'long' condition. Similarly, against our prediction, participants were again more likely to select the outcome-congruent response when the final block position was 2 or 3, 73% [60%, 88%], but the effect was again 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 the finish line was moved forward such that only position 4 crossed it (but not position 3). When the ground truth position was 3, participants were more likely to select the outcome-congruent position 2 than position 4 (59% [46%, 72%]) but, against what we predicted, the credible interval of the estimate did not exclude 50%. As predicted, participants' accuracy for position 4 (81% [71%, 90%]) was greater than that for the other three positions (51% [41%, 62%], $B = 1.44$ [0.86, 2.01]).

[1] We pre-registered Bayesian analyses for Experiment 2 because it was easier to implement ordinal mixed effects regression models this way.

Discussion

Experiment 2 again shows a pattern of results consistent with the idea that participants built a goal-contingent abstraction of the task. Like in Experiment 1, participants were able to learn the causally relevant information for predicting the outcome. A consequence of building this representation was that participants produced systematic errors when confronted with a 'surprise test' for which their learned task representation was inadequate. Participants were more likely to recall incorrect outcomes that were consistent with the causal rule that they had learned. Participants produced these errors even though they had ample experience with the setting. Indeed, in the 'long' condition and 'conjunctive' condition, participants viewed each of the four clips four times in the prediction task.

Unlike what we predicted, and unlike what we found in Experiment 1, participants made systematic errors even when they had little training experience in the 'short' condition. This suggests 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%. In all three conditions, participants were very unlikely to select a response that was inconsistent with the outcome, such as selecting position 3 when the correct position was 2 in the 'long' or 'short' conditions. This suggests that participants may have paid particular attention when the outcome was close to the finish line.

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 nicely illustrates how people learned causal abstractions that were suited to the task at hand.

Limitations and future directions

How do people build mental models of the world? Our paper adds to the existing literature suggesting that whatever shape these models take, they are constructed to suit the task at hand. The errors people make reveal that they build abstract models that privilege causally relevant information. Of course, our study is just a first step in prodding the abstract causal models in people's minds. To uncover more, we will likely need a suite of tools. We can ask people to draw (Huey, Walker, & Fan, 2021), or to recall (Ho et al., 2022). We can look at their eyes (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017), or look at their brains (Muhle-Karbe et al., 2023). And we can develop computational models of resource-rational agents that tell us where to look (Bates et al., 2019).

Acknowledgments

TG was supported by a research grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI). We thank David Rose for feedback on the manuscript.

References

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

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

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

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

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

Beller, A., Bennett, E., & Gerstenberg, T. (2020). The language of causation. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3133–3139). Cognitive Science Society.

Beller, A., Xu, Y., Linderman, S., & Gerstenberg, T. (2022). Looking into the past: Eye-tracking mental simulation in physical inference. *Cognitive Science Proceedings*.

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.

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

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72.

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.

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, 104842.

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

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

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

Halpern, J. Y., & Hitchcock, C. (2011). Actual causation and the art of modeling. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (pp. 316–328). College Publications.

Ho, M. K. (2019). The value of abstraction. *Current Opinion in Behavioral Sciences*, *29*, 111–116.

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.

Huey, H., Walker, C. M., & Fan, J. E. (2021). How do the semantic properties of visual explanations guide causal inference? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

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

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, 0956797620957610.

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

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

Smith, K. A., & Vul, E. (2012). Sources of uncertainty in intuitive physics. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

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.

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.

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).

Woodward, J. (2015). The problem of variable choice. *Synthese*, *193*(4), 1047–1072.

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

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

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

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2022). Mental Jenga: A counterfactual simulation model of causal judgments about physical support. *PsyArXiv*.