Looking into the past
Eye-tracking mental simulation in physical inference

Ari Beller (beller@stanford.edu), Department of Psychology, Stanford University

Yingchen Xu (yingchen.xu.21@ucl.ac.uk), Department of Computer Science, University College London

Scott Linderman (scott.linderman@stanford.edu), Department of Statistics, Stanford University

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

Abstract

Mental simulation is a powerful cognitive capacity that underlies people's ability to draw inferences about what happened in the past from the present. Recent work suggests that eye-tracking can be used as a window through which one can study the process of mental simulation in intuitive physics tasks. In our experiment, participants have to figure out in which of three holes a ball was dropped in a virtual Plinko box. We develop a computational model of human intuitive physical reasoning in Plinko that runs repeated simulations in a noisy physics simulator in order to infer in which hole the ball was dropped. We evaluate our model's behavior against multiple human data signals: trial judgments, response times, and eye-movement data. We find that a model that sequentially samples simulations while balancing uncertainty and reward best explains the patterns of participant behavior we observe in these three signals.

Keywords: mental simulation; intuitive physics; causal inference; eye-tracking; computational modeling.

Introduction

Imagine walking into your dining room and noticing one of your favorite vases shattered on the floor. Your eyes quickly flit up to its former location on the dinging room table, and you spot your mischievous cat, Whiskers, looking guilty. Without a moment's hesitation an explanation for what happened pops into your head. Whiskers was playing where he wasn't supposed to, bumped the vase, and gravity and physics did the rest.

This seemingly unremarkable sequence of thoughts actually exhibits the components of an impressive cognitive processing capacity. Having observed an unexplained outcome, you were able to utilize your intuitive knowledge of how the world works to imagine a plausible story that explains the data you observed. This ability to infer past causes from present events is constantly at work in human thought. It comes out in relatively mundane interactions with our rambunctious cats, but also in more complicated settings where people must reconstruct the past from the present like a detective determining what happened at a crime scene.

How do people perform these impressive feats of inference? Prior research suggests that intuitive theories encoding rich causal knowledge about the structure of the world can support these powerful leaps of reasoning backward from observed effects to latent causes (Gerstenberg & Tenenbaum, 2017; Lake, Ullman, Tenenbaum, & Gershman, 2017; Wellman & Gelman, 1992). Work in intuitive physics in particular has highlighted the role of mental simulation as a cognitive mechanism supporting probabilistic inference about possible physical histories (e.g. Smith & Vul, 2014). Building on this hypothesis, a modeling tradition has emerged over the past ten years that uses approximate physics engines to explore how mental simulation can support a wide variety of intuitive physical inferences (Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). Though these models have certain limitations as a full description of human physical reasoning (Ludwin-Peery, Bramley, Davis, & Gureckis, 2021), they provide a rich computational tool set that allows cognitive psychologists to propose explicit hypotheses yielding quantitative predictions, and compare those predictions against human behavioral data.

In concert with these developments in modeling physical inference, new methods have been developed for extracting behavioral signals of human physical thought. Eye-tracking in particular has proven a promising approach. In a variety of intuitive physical tasks, researchers have captured human eye-data to investigate claims about mental simulation (Ahuja & Sheinberg, 2019; Crespi, Robino, Silva, & de'Sperati, 2012; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017). Eye-data yield a moment-to-moment trace of human behavior throughout the process of making a physical judgment, augmenting standard behavioral measures and providing rich empirical fodder for making inferences about human cognition.

In this study, we work to bring together modeling tools for intuitive physics and eye-tracking. We examine participant behavior in Plinko, an intuitive physics task developed by Gerstenberg, Siegel, and Tenenbaum (2021). In their study, participants performed either a prediction task or an inference task. Here we focus on the inference task which is illustrated in Figure 1. Participants were presented with images showing the final location of the ball and asked to infer in which hole the ball was dropped. Gerstenberg, Siegel, and Tenenbaum found that a model that relies on physical simulation outperformed alternatives that only used heuristic cues, suggesting that mental simulation is likely at play in participants' inferences in this task. However, their initial study only considered human judgment data. Here, we augment the Plinko paradigm with eye-tracking data as well as response time data

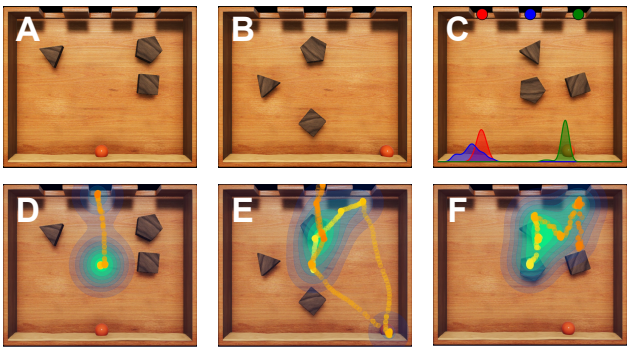

Figure 1: A–C: Sample stimuli from the Plinko task. Participants are presented with still images like those on the top row, and asked to judge which hole they think the ball most likely fell from. Panel C also shows kernel density estimates from the uniform sampler computed from multiple simulations from each hole. D–F: Sample traces of individual participants' eye-movements for each of the stimuli above. The colored dots represent eye-positions over time where yellow dots are closer to the beginning of the trial and orange dots are closer to the end. The green-blue density reflects the amount of time spent foveating in that location.

to better understand the underlying cognitive processes that support causal inferences in this task. In particular, these additional data sources allow us to provide stronger evidence for the role of mental simulation in causal inference.

The paper is organized as follows. We begin by describing our modeling framework. We present noisy physics simulation as a tool for modeling human intuitive physical thinking in our domain. We then describe the uniform sampling model first introduced by Gerstenberg, Siegel, and Tenenbaum (2021) as a model of human judgment in the Plinko task. We proceed to introduce a sequential sampling model, that builds on this prior approach to better characterize the cognitive process at work in the Plinko task. We then introduce the task and discuss how well the different models account for the human data we collected. We highlight how sequential sampling helps us better explain participant behavior, capturing a strong trend in participant response time and skewed distributions of participant eye-movement. We close with a brief consideration of future directions.

Modeling causal inference

Our inference models for Plinko are built on an approach that uses noisy physics simulators as a model for human intuitive physical thought. We describe the approach here, and then present two models that utilize mental physical simulation to perform inference in the Plinko task.

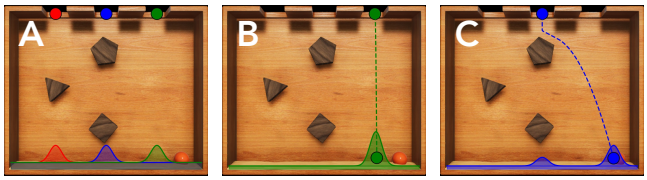

Figure 2: A sequence of behavior of the sequential sampler. Panel A shows the initial conditional distributions for each hole. The model initially favors the hole closest to where the ball is (the green hole). In panel B the model simulates a ball drop from this hole and updates its kernel density estimate. However doing so actually decreases the model's belief that the ball was dropped in that hole (it also decreases the entropy). In panel C, the model proceeds to consider the next best hypothesis and simulates a drop from the blue hole.

Modeling physical reasoning through mental simulation

Following Gerstenberg, Siegel, and Tenenbaum (2021), we model physical inference by running repeated simulations in a noisy physics simulator. This approach builds on a broader literature that uses noisy physics engines as a model of human intuitive physical reasoning (Battaglia et al., 2013; Gerstenberg, Goodman, et al., 2021; Ullman et al., 2017). In general, people cannot perform exact mental physical simulation, and noisy simulators allow us to capture that uncertainty embedded in the cognitive process. Uncertainty in mental physical simulations arises from many sources (Smith & Vul, 2012). We model physical uncertainty in Plinko using two sources of noise: 1) uncertainty in the angle at which the ball is dropped from a hole ("drop noise"), and 2) uncertainty in the ball's magnitude of velocity after it collides with an obstacle or the wall ("collision noise"). Gerstenberg, Siegel, and Tenenbaum (2021) showed that these two sources of uncertainty were sufficient for accurately capturing participants' predictions and inferences in Plinko. "Drop noise" is implemented by adding Gaussian noise to the true angle of the drop, while "collision noise" is implemented by multiplying the ball's true exit velocity with a value generated from a Gaussian distribution. The parameters for these distributions are set to the same values used by Gerstenberg, Siegel, and Tenenbaum (2021) in their original study: $\mathcal{N}_{\text{drop noise}}(0, 0.2)$ and $\mathcal{N}_{\text{collision noise}}(0.8, 0.2)$. The mean value of "collision noise" is below 1 to capture participants' systematic tendency to underestimate how far the ball will bounce off the obstacles.

To investigate more precisely whether, and if so, how participants use mental simulations to perform inference in Plinko, we develop two computational models. The first simulates uniformly from each of the three holes, re-implementing the same computational model developed by Gerstenberg, Siegel, and Tenenbaum (2021). Gerstenberg, Siegel, and Tenenbaum found that this uniform sampler did a good job of capturing patterns of participant judgments in

Plinko. However, we see this model as implausible as a model of the underlying cognitive process. Prior work suggests that, when making physical judgments, humans consider and act on certain hypotheses preferentially rather than considering all possibilities uniformly (Dasgupta, Smith, Schulz, Tenenbaum, & Gershman, 2018). We expect that here too participants are more focused on the plausible drops that could have given rise to the observed outcome. To sharpen this intuition, we design a sequential sampling model that iteratively simulates from hypotheses that balance reward and uncertainty.

Uniform sampling model

The uniform sampling model determines the most probable hole by performing Bayesian inference through repeated simulation in a noisy physics engine. Our uniform sampler computes a posterior distribution on holes h_i given the observed final location of the ball x_{obs} according to Bayes' Rule:

$$p(h_i|x_{obs}) \propto p(x_{obs}|h_i)p(h_i) \quad (1)$$

Here the prior on holes $p(h_i)$ is assumed to be uniform. The model estimates the likelihood $p(x_{obs}|h_i)$ by simulating a fixed number of samples from each of the holes, and computing a kernel density estimate for each hole with the samples dropped from that hole. The model computes the likelihood of x_{obs} using the kernel density estimate from each hole. For example, Figure 1C shows the kernel density estimates computed after taking a fixed number of samples from each of the holes. The location of the ball has a high probability under the green distribution so the likelihood is high, but under the other two hypotheses the likelihood is very low. The uniform sampler has two free parameters, the number of samples dropped from each hole and the bandwidth of the Gaussian kernel that's used to generate the kernel density estimates.

Sequential sampling model

While the uniform sampling model from Gerstenberg, Siegel, and Tenenbaum (2021) did a good job of capturing participants' inferences about where the ball fell from, we think it is implausible as a model of the underlying cognitive processes for two reasons. First, the model performs a fixed number of simulations on every trial. However, some trials may be easier to assess than others. Second, the model pays equal attention to all three holes. However, it's likely that some holes strike participants as better candidates than others (see Figure 1). To account for these intuitions, we design a sequential sampling model that iteratively determines whether to sample another hole and if so which one.

We formulate the process of simulation allocation in Plinko as an explore-exploit tradeoff (Schulz et al., 2019). Participants want to find a good hypothesis, and repeated simulation (exploitation) from a particular hole could help them increase their confidence that they have found the correct one. At the same time, sampling from a particular hypothesis to the exclusion of other possibilities could prevent them from finding a better alternative hypothesis. Thus, participants are also motivated to explore different possibilities and see whether there are good options that they might be missing.

In order to formalize a decision-making agent to model these dueling pressures, we cast sequential simulation choice in Plinko as a multi-armed bandit, a classic paradigm for balancing exploration and exploitation in sequential decision making tasks (Slivkins, 2019). We implement an upper confidence bound algorithm to solve the bandit problem posed by our setting. The upper confidence bound agent decides which action to take based on a weighted combination of the expected reward and uncertainty of each action (Lai & Robbins, 1985).

Figure 2 provides a graphical illustration of the sequential sampler's behavior. The model initializes in panel A with a prior expectation that the ball will fall close to the hole it was dropped. It proceeds in panel B to consider the hole closest to the ball, the green hole. The model simulates from this hole and finds that doing so actually makes this hypothesis less appealing. The model is now more certain that the green hole would not have given rise to the observed outcome. The model proceeds in panel C to consider the blue hole, and finds this hypothesis more plausible.

Formally, our model chooses to simulate from a hole h_i that maximizes the following utility function:

$$U(h_i; x_{obs}) = p(x_{obs}|h_i) - \omega \int p(x|h_i) \log p(x|h_i) dx, \quad (2)$$

$p(x_{obs}|h_i)$ represents the reward term. The model preferentially simulates from holes that assign high probability to the observed location of the ball. $-\int p(x|h_i) \log p(x|h_i) dx$, the entropy of the conditional distribution, is the uncertainty term. If the conditional distribution for a hole is very wide and the model is uncertain about where the ball will fall when dropped from that hole, the model is incentivized to simulate from that hole to reduce its uncertainty. The free parameter ω tunes the balance between these two incentives.

On each trial, we initialize the kernel density for each hole with a small Gaussian bump under the location of the hole, reflecting a bias to expect a priori that the ball will fall closer to the hole. The model proceeds to choose hypotheses sequentially according to this utility function, simulate from the corresponding holes, and update its kernel density estimate for that hole with the value of the new sample. The sample weight, a free parameter of the model, determines how strongly each simulation affects the shape of the conditional distribution. After each simulation, the model computes a posterior distribution on holes by re-normalizing the likelihoods and checks whether the entropy of that distribution has fallen below a decision threshold (another free parameter). When it does, the model terminates the sampling procedure and selects the hole with the highest probability under the posterior.

All together, the model has four free parameters: the decision threshold, the reward-uncertainty tradeoff, the bandwidth of the kernel density, and the sample weight.

Experiment

We measure participant behavior in the Plinko inference task. Participants saw images showing the ball at the bottom of the Plinko box (see Figure 1). Participants' task was to guess in which hole the ball was dropped. In addition to this judgment data, we collected response time and eye-movement data as well. The materials, data, model and analyses scripts can be accessed here: `https://github.com/cicl-stanford/tracking_inference`

Methods

Participants We recruited 30 participants through Stanford's community recruitment platform and undergraduate student credit pool (*age*: mean = 25, sd = 8; *gender*: 14 female, 16 male; *race*: 12 Asian, 15 White or Caucasian, 3 unclear racial categories). Community members were compensated at a rate of $11 per hour and students were compensated with course credit.

Design There are three obstacles in each Plinko box trial: a triangle, a square, and a pentagon. Each obstacle can occupy one of nine possible locations that form a grid evenly spaced under each of the three holes. Trials are generated by randomly sampling a location for each obstacle from this grid with the constraint that no obstacle can occupy the same location. Once it has a location, each obstacle is randomly offset and randomly rotated. Once the obstacles are fixed, a single simulation is run from a random hole in a deterministic physics engine to determine the resting place of the ball.

After generating a large number stimuli, a subset were selected as trials. The difficulty of the inference was varied across the selected set, where difficulty here is defined as the entropy of the posterior distribution under the uniform sampler model. Trials were further selected to tease apart participant judgments in experimental conditions where participants would have access to additional auditory cues as well as visual information. However in this work, we focus on the experimental condition in which participants received only visual evidence.

Procedure Participants were first presented with six training videos depicting drops from each of the holes to orient them to how the physics works. We then instructed them that on each trial they will view static stimuli depicting the end-state of a drop, and that their task would be to infer which hole they think the ball fell from.

We proceeded to calibrate the eye-tracker for each participant. We tracked each participant's right eye using an SR Research Eyelink 1000 sampling at 1000 Hz. Participants rested their head on a chin rest fixed 54 cm from the display.

In the main stage of the experiment, participants performed the inference task on 152 trials, two of which were training trials. After training, the order of test trials was randomized between participants. Before each trial, participants fixated at the center of the screen to initiate the trial. The trial terminated when the participant pressed 1, 2, or 3 on the keyboard indicating their judgment that the ball had dropped from the corresponding hole. We recorded participants' judgments, response times (from stimulus onset to the keyboard response), and eye-movements. Every 30 trials we had a break period for the participant to rest their eyes and remove their head from the chin rest. After two break periods we re-calibrated the eye-tracker.

Results

To evaluate model performance, we first optimized the parameters of the uniform and sequential sampling models with a grid search. For each data signal, we computed the squared error between the model prediction and the human data as a measure of performance. We chose the model parameters that minimized the average rank in accounting for the three data signals. We fit model parameters on half our participants ($n = 15$) and report model performance on the complete set.

Judgments Participant judgments present the most straightforward test of our model's behavior. On a given trial, the model should select holes that match the distribution of participant responses. For the uniform sampler, this comparison is simple. The model directly produces a posterior that can be compared to the participant distribution. To produce a distribution of behavior for the sequential sampler, we run the model multiple times and compute the proportion of runs in which the sequential sampler selected each hole.

The scatter plots in Figure 3A illustrate model performance on the judgment task. On the y-axis is the proportion of participants that selected a particular hole on a given trial, and on the x-axis is the proportion of model runs where a particular hole was selected for the sequential sampler and the posterior estimate for the uniform sampler. Each point represents the judgment for a particular hole on a particular trial. Both models correlate strongly with participant judgments, though this effect is driven in large part by clusters of model judgments at 0 and 1. 358 of 450 judgments are either 0 or 1 for the sequential sampler and 187 of 450 judgments for the uniform sampler are either 0 or 1. There remains substantial residual variance between the two extremes. Overall the two models' performance is similar in terms of how well they correlate with participants' responses, as well as the average error between model predictions and participant judgments. The two models correlate strongly with each other as well ($r = 0.90$). Digging deeper into our additional data signals is important to pull apart their behavior.

Response Times The time it takes participants to respond differs across trials. On some trials participants are fast (e.g., Figure 1D), whereas on others, it takes them longer to judge in which hole the ball was dropped (e.g., Figure 1E). To evaluate whether our models can explain the time it takes for participants to figure out what happened, we compute a measure of model response time for each trial and compare that value to our human data. Specifically, each time we run the model on a given trial we count the number of collisions that take place across all the simulations that the model ran. This gives

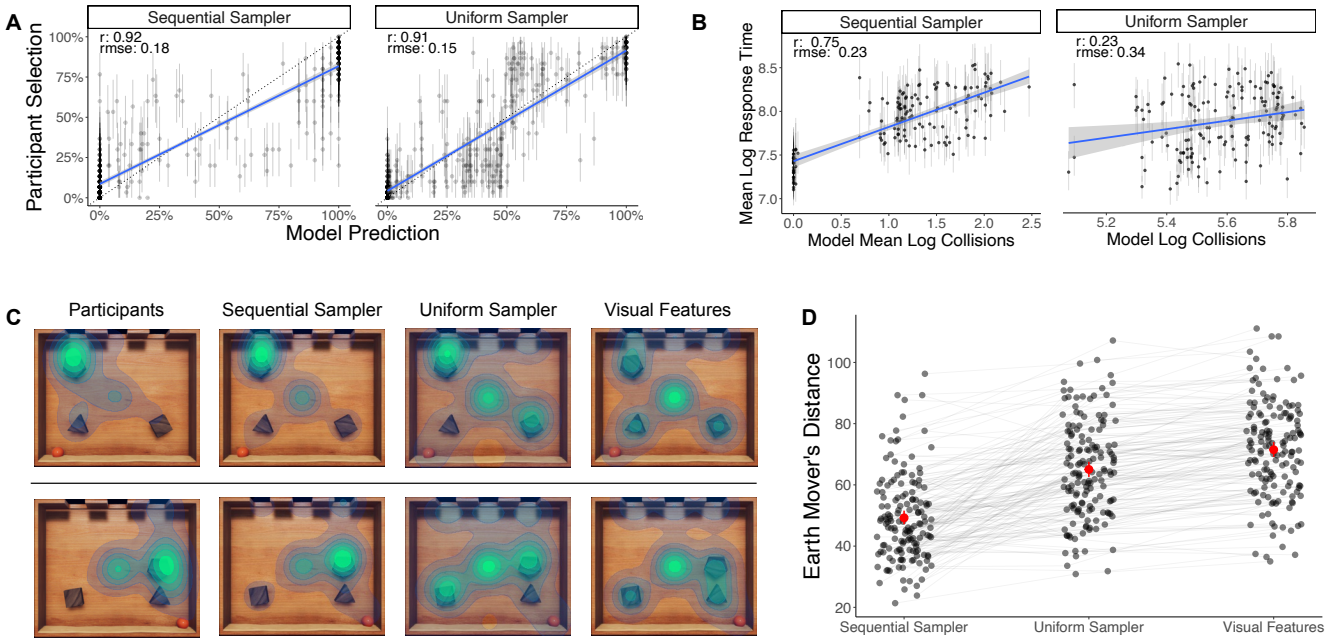

Figure 3: Model comparison between the sequential sampler and the uniform sampler. **A: Judgments** Scatter plots of model prediction against human judgments. The x-axis depicts the model estimate for each hole, and the y-axis depicts the proportion of participants who selected that hole. Each point represents a particular hole on a particular trial (150 trials $\times$ 3 holes = 450 data points). **B: Response Times** Scatter plots comparing model predictions to human response times for the 150 trials. To predict participants' response times, the sequential sampler uses the mean log collisions across its multiple runs, and the uniform sampler uses the log total number of collisions from its single run. Each point represents a single trial. **C: Eye Data** Heatmaps of participant eye-movement and model predictions on two sample trials. **D: Eye Data** Comparison between participant heatmaps and model heatmaps on each trial using earth mover's distance as a measure (lower is better here). The visual features baseline produces a heatmap using only visual features of the scene including the position of the obstacles, the holes, and where the ball landed.

us a sense of how much "thought" our model put into figuring out what happened. Counting the number of collisions rather than the number of simulations accounts for the fact that certain simulations may be more complicated and require more cognitive effort than others. Human response time data has a characteristic long tail. While most trials are performed relatively quickly, a small number of trials take a long time for participants to respond. Assessing model fit with a regression would be strongly affected by outliers. To mitigate this effect, we log-transform the human response time data and compute the mean log response time for each trial.

The sequential sampler response time measure exhibits the same characteristic long tail, so we prepare the model response time measure in an analogous way. We log transform the collision count on each run, and for each trial we compute the mean log collision count as a measure of model response time. For the uniform sampler, the number of simulations is fixed, and thus the number of collisions varies minimally when the model is run multiple times. As such, running the uniform sampler a single time gives us a relatively good estimate of the mean if we were to run it many times. In comparing this model to human response times, we take the log number of collisions from a single run.

The scatter plots in Figure 3B illustrate how well each model accounts for participants' average response times. Here the difference between the two models is substantial. The sequential sampler explains much more of the variance in human response times compared to the uniform sampler. For the sequential sampler, part of what's driving the higher correlation is that it accurately captures that participants respond quickly for some of the trials. These are cases such as the one shown in Figure 1A where both the sequential sampler and participants quickly figure out that the ball was dropped in hole 2. Even if this cluster of trials is removed, the sequential sampler still correlates better with human response times ($r = 0.37$) than the uniform sampler does.

Eye Data The eye-data is a complex data signal that unfolds across time in different types of movements (fixations and saccades). For this analysis, we simplify the complexity of this signal by looking at the eye-data aggregated across participants and time. Within each trial, we compute a two-dimensional density estimate with a Gaussian kernel based on all the eye-data samples collected from all participants on that trial. This kernel density provides us with a description of the

distribution of participant gaze locations on each trial. The distributions of participant eye-data give an activation value at each pixel of the Plinko box. In the sample cases in Figure 3, deep green areas represent locations with high activation values, while more faded blue areas have relatively lower values. Areas without color have activation values near zero.

To predict where participants are looking, we define a set of feature maps that are derived from the model behavior. For example, a key feature that we believe is important in explaining the patterns of participant looking behavior is the location of the collisions that the model simulates when considering how the ball would fall from a given hole. Both the sequential and the uniform sampler produce a set of collisions on each trial, and we can use the locations of these collisions as samples to compute a kernel density estimate, just as we did for participants' eye data. The resulting feature map has an activation at each pixel that then serves as a predictor for the corresponding activation in the human distribution. We can compute multiple feature maps of this type and then use a regression to determine a weighted combination of features that best explains the pattern of participant eye-data.

We compute feature maps for four *dynamic features*: the locations of the simulated obstacle collisions, the locations of the simulation drops, the locations of the simulated wall collisions, and the locations of the simulated collision with the ground. We also compute feature maps for four *visual features*: the locations of the obstacles, the locations of the holes, the location of the ball, and the center of the Plinko box. As a baseline to compare against our two simulation models, we compute a regression from the visual features to the participant data with no dynamic features included. The performance of this 'visual features' model relative to the simulation models shows whether dynamic features are important for predicting human eye-gaze.

We measure the difference between the distribution of participant eye-movements and the predicted distribution of fixations for each of our models using the earth mover's distance (Rubner, Tomasi, & Guibas, 2000). The results of this comparison are illustrated in Figure 3D. On the left we see distances from the sequential sampler distributions to the participant distributions for each trial (Mean: 49.23, 95% confidence interval: [46.93, 51.54]), in the middle the analogous distances for the uniform sampler (Mean: 65.05 [62.60, 67.65], and on the right our visual feature baseline (Mean: 71.49 [69.08, 73.94]). The sequential sampler outperforms both alternatives on this measure, while the uniform sampler is only somewhat better than the visual feature baseline.

The sample trials in Figure 3C give a sense of why the sequential sampler outperforms the uniform sampler. Looking to the participant data on the left, we see that the distributions are notably skewed toward plausible hypotheses that could have given rise to the observed outcome. While the sequential sampler is able to accommodate these patterns, the uniform sampler struggles to do so. It pays a substantial amount of attention to collision points that participants altogether ignore (such as the collisions with the square that result from the ball being dropped in hole 3 in the top example, or in hole 1 in the bottom example).

General Discussion

In this paper we looked at the role that mental simulation plays in how people make causal inferences about what happened in the past. We used the Plinko task developed by Gerstenberg, Siegel, and Tenenbaum (2021). While, Gerstenberg, Siegel, and Tenenbaum only collected human judgments, we also assessed response times, and eye-movements to gain further insights into the underlying cognitive processes that support causal inferences in this task.

We compared participants' behavior in the task with the predictions of two computational models. The uniform sampler (which runs the same number of simulations from each hole) was able to capture participants' judgments. However, it didn't account well for participants' response times, or their eye-movements. In contrast, the sequential sampler did a better job of capturing all three data signals. The sequential sampler runs simulations one by one, whereby its choice of what simulation to run next is guided by a trade-off between maximizing reward (i.e. simulating drops from holes where the ball ends up close to where it actually was) and minimizing uncertainty (i.e. considering holes it hasn't explored before).

The sequential sampler makes a first step toward a more complete process model of how people arrive at their inferences about what happened. In the future, we would like to capture not only aggregated eye-data but also the specific eye-movements that participants produce (including fixations and saccades). Currently, the eye-data that our models produce don't really play a causal role in the inference. However, we believe that people use their eye-movements to systematically reduce perceptual uncertainty about the position and rotation of the obstacles, as well as dynamic uncertainty about how the ball would fall from a hole. Building a model which represents the time-course of participant eye-movements may unlock further insights into the cognitive process at play.

Another direction that merits further investigation is how participants go beyond visual evidence to figure out what happened. Gerstenberg, Siegel, and Tenenbaum (2021) explored how participants combined visual and auditory evidence in their causal inferences. Participants first heard the sounds that the ball made when it was dropped into an occluded box. The cover was then revealed so that participants saw the final position of the ball. It will be interesting to explore how this auditory information affects the process by which people are considering different hypotheses over time. For example, it is plausible that participants may use the sounds to quickly rule out certain hypothesis ("I heard a collision so the hole must have an obstacle underneath."), and then rely on more detailed mental simulations to differentiate between the remaining hypotheses. We look forward to studying more how people look into the past!

Acknowledgments

Scott Linderman and Tobias Gerstenberg were supported by a seed grant from Stanford's Human-Centered Artificial Intelligence Institute (HAI). The experiment was approved by Stanford's Institutional Review Board.

References

Ahuja, A., & Sheinberg, D. L. (2019). Behavioral and oculomotor evidence for visual simulation of object movement. *Journal of vision*, *19*(6), 13–13.

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.

Crespi, S., Robino, C., Silva, O., & de'Sperati, C. (2012). Spotting expertise in the eyes: Billiards knowledge as revealed by gaze shifts in a dynamic visual prediction task. *Journal of Vision*, *12*(11), 1–19.

Dasgupta, I., Smith, K. A., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2018). Learning to act by integrating mental simulations and physical experiments. *bioRxiv*.

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., Siegel, M. H., & Tenenbaum, J. B. (2021). What happened? reconstructing the past from vision and sound. *PsyArXiv*.

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

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, *6*(1), 4–22.

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.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, *127*, 101396.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, *40*(2), 99–121.

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, *116*(28), 13903–13908.

Slivkins, A. (2019). Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*.

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.

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.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337–375.