Multimodal inference through mental simulation

Ari Beller[1], Yingchen Xu[2], Max Siegel[3], Satchel Grant[1], Alan Brown[1], Jan-Philipp Fränken[1], Joshua B. Tenenbaum[3], Scott W. Linderman[1], and Tobias Gerstenberg[*1]

[1]Stanford University

[2]University College London

[3]Massachusetts Institute of Technology

Author Note

[*]Corresponding author: Tobias Gerstenberg https://orcid.org/0000-0002-9162-0779; gerstenberg@stanford.edu. All code and materials available at `https://github.com/cicl-stanford/multimodal_plinko`.

We would like to thank the members of the Causality in Cognition lab for their many helpful comments in the preparation of this work. TG was supported by a grants from Stanford's Human-Centered Artificial Intelligence Institute (HAI) and Cooperative AI. SWL was supported by fellowships from the Sloan and McKnight Foundations.

Data from Experiment 1 have appeared previously as a pre-print in Gerstenberg, Siegel, and Tenenbaum (2018). Data from Experiment 2 have appeared in the *Proceedings of the Annual Conference of the Cognitive Science Society* in Beller, Xu, Linderman, and Gerstenberg (2022).

Abstract

The ability to infer the past from what we perceive in the present is a key capacity of human cognition. Witnessing a broken vase, humans will automatically bring to mind a causal story of what happened. Multiple sources of sensory evidence can support this inference. Seeing the broken pottery tells you something, but hearing the crash tells you even more. In this work, we explore people's inferences about the past from multimodal evidence. We present a physical reasoning paradigm called Plinko. In the prediction task, participants must determine where a ball that is dropped into a box with obstacles will land. A computational model that uses mental simulation in an Intuitive Physics Engine captures participant predictions very well. In the inference task, participants must infer which hole in the box the ball fell from. Across conditions, participants are presented with different combinations of visual and auditory cues, and must combine this information to determine what happened. We develop a sequential sampling model that selectively simulates from promising hypotheses, and demonstrate that this model accurately captures participants' judgments and eye-movements. By coordinating sensory evidence in an underlying causal representation of the physical world, this simulation approach is able to capture complex multimodal inferences that go beyond traditional approaches to multimodal integration.

Keywords: multimodal integration; mental simulation; intuitive physics; eye-tracking; Bayesian inference

Multimodal inference through mental simulation

Introduction

Imagine you are coming home from work and as you walk in the door you hear a crash from the dining room. You rush over to see what happened and immediately spot your favorite vase shattered on the floor. Your eyes quickly flit to its former location on the dining room table, and there sits your cat, Whiskers, looking guilty. In a flash, an explanation for what happened pops into your head. Whiskers was playing where he didn't belong, bumped the vase, and gravity did the rest.

This sequence of thoughts might seem mundane, but it exhibits the components of an impressive cognitive capacity. Having observed a surprising situation, you were able to combine disparate signals, sounds and sights, to construct a plausible story of what happened. This ability is foundational to human cognition. Every day, we are deluged by a sea of sensory information, and we must sift through this complex multimodal experience to extract reliable causal explanations that allow us to understand our environment. In low-stakes interactions with mischievous cats, but also in high-stakes situations like a detective reconstructing what happened at the scene of a crime, we are constantly interpreting our sensory experience to uncover these causal stories that go beyond what we can directly perceive (Chen & Scholl, 2016). Understanding how people achieve this cognitive feat is a central question in psychological science.

The ability to integrate multimodal information is a critical piece of the puzzle. Coordinating information across sensory channels limits the number of possible situations that could explain a given experience (Alais & Burr, 2019; Stein & Meredith, 1993). Psychophysics research has demonstrated that humans are quite adept at combining different modalities of sense data. In perceptual tasks, such as spatial localization and size discrimination, people optimally combine evidence from multiple sensory signals, weighing the different modalities in accordance with their reliability (Alais & Burr, 2004; Ernst & Banks, 2002; Jacobs, 1999; Wozny, Beierholm, & Shams, 2008). However, concurrent

sensory cues don't always come from the same source. Inferences about the latent causes of sensory experience shapes how and when people integrate evidence across different modalities (Körding et al., 2007; Rohe & Noppeney, 2015). The interpretation of multimodal evidence takes place in the context of a continuous inferential process aimed at ascertaining these latent causes of experience (Shams, 2012; Shams & Beierholm, 2010). But while traditional multimodal perception paradigms present stimuli with a simple causal setup, daily sensory perception unfolds against the backdrop of a richly structured causal environment (Kim & Schachner, 2025).

To navigate this complexity, humans bring to bear a powerful set of cognitive tools. Over development, humans fine-tune intuitive theories, structured causal models representing how objects and agents relate to and influence one another (Gerstenberg & Tenenbaum, 2017; Lake, Ullman, Tenenbaum, & Gershman, 2016; Wellman & Gelman, 1992). Intuitive theories allow people to simulate the causal dynamics of their environment, supporting prediction of unseen possible futures and backward inference to the causes of unobserved pasts (Hegarty, 2004). Intuitive theories can also explain how an underlying causal process gives rise to disparate sensory signals, unifying multimodal evidence in a common representation (Erdogan, Yildirim, & Jacobs, 2015; Yildirim & Jacobs, 2014; Ying et al., 2025). In light of their importance for understanding how people perform inferences across wide-ranging domains, a modeling tradition has emerged in cognitive science that aims to reverse-engineer the nature of intuitive theories (Battaglia, Hamrick, & Tenenbaum, 2013; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). For example, in the domain of intuitive physics, researchers have developed computational models to explain human behavior in a wide variety of tasks including physical prediction (Hamrick, Smith, Griffiths, & Vul, 2015; Smith, Dechter, Tenenbaum, & Vul, 2013; Smith & Vul, 2013), judgments of physical support (Battaglia et al., 2013; Zhou, Smith, Tenenbaum, & Gerstenberg, in press), causal judgments in physical scenes (Beller & Gerstenberg, 2025; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021), perception of occluded objects

(K. W. Wong, Bi, Soltani, Yildirim, & Scholl, 2023), and human intuitions of liquid flow (Bates, Yildirim, Tenenbaum, & Battaglia, 2019). Moreover, eye-tracking has emerged as a promising empirical method for indexing mental simulation in physical reasoning tasks (Ahuja et al., 2024; Ahuja & Sheinberg, 2019; Crespi, Robino, Silva, & de'Sperati, 2012; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017). Gaze patterns provide a continuous trace of participants' attention, revealing what they are interested in and suggesting how they use mental simulation to make judgments about physical scenes.

In this work, we examine how people integrate evidence from multiple sensory modalities in an underlying causal model of the world in order to perform complex physical inferences. We present a novel task domain: Plinko – a flexible experimental setting for assessing participant reasoning in physical prediction and inference tasks. We develop a model of multimodal inference that uses a noisy physics simulator to approximate the cognitive processes that people engage when reasoning in this domain. In Experiment 1, we demonstrate that this model achieves a very high level of predictive accuracy in explaining participant predictions and outperforms alternatives, suggesting that mental physical simulation is a key mechanism supporting physical reasoning in the Plinko domain. In Experiment 2, we demonstrate how this model explains the coordination of sensory cues from multiple modalities, capturing participant judgments and eye-movements in our task, and thereby deepening our understanding of how people use their causal knowledge to navigate their complex multimodal experience.

Results

Plinko Domain

The Plinko Domain is a physical reasoning setting that allows us to examine how participants make judgments about their physical environment (see Figure 1). The Plinko box has three holes from which a ball can be dropped. During a drop, the ball bounces off obstacles as it falls to the floor, creating auditory cues as it collides with the various obstacles and with the walls. This setting supports different physical reasoning tasks.

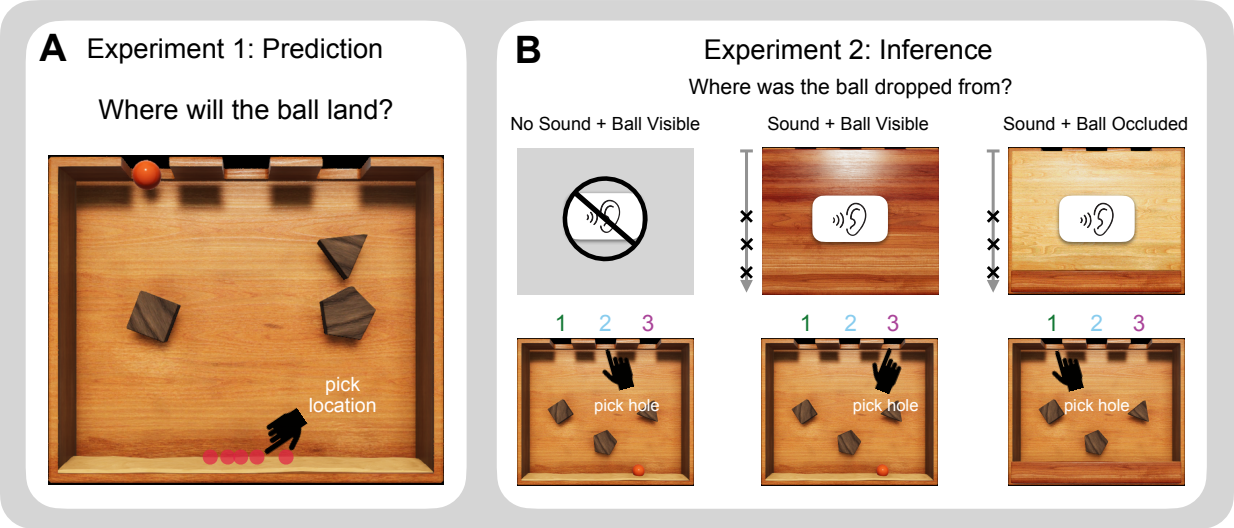

Figure 1

Illustration of the Plinko Domain. Plinko is an intuitive physics setting designed for the study of diagnostic inference from multimodal sensory evidence. ***A*** *In the prediction task, participants see the drop location and must predict where the ball will land, clicking multiple times to indicate potential uncertainty.* ***B*** *In the inference task, participants receive combinations of visual and auditory cues and must infer where the ball fell from. In the 'no sound + ball visible' condition, participants only see the final ball location. In the 'sound + ball visible' condition, participants first hear the sounds of the ball falling (while the box is covered), and then see the final ball location. Collision sounds are indicated by the* $\boldsymbol{x}$ *points on the vertical timeline. In the 'sound + ball occluded' condition, participants again hear the sounds of the ball falling, but the final ball location is covered. They can only see the obstacles.*

Figure 1a illustrates the prediction task. Participants see which hole the ball will be dropped from and are asked to predict its endpoint. Participants provide ten clicks indicating the location along the floor of the Plinko box that they think the ball will land. This response format allows participants to express uncertainty. If a participant is very confident about where the ball will land, they can click many times in the same spot. If they are more uncertain and believe a broader range of outcomes are possible, they can indicate that with multiple clicks in different locations.

Figure 1b illustrates the inference task. In this task, participants receive different forms of visual and auditory evidence and must infer which hole the ball fell from. We

examine participant inferences in three different conditions. In the 'no sound + ball visible' condition, participants only see an image that shows where the ball ended up. In the 'sound + ball visible' condition, participants are first presented with a covered Plinko box and hear the sounds that the ball makes as it's dropped. Participants hear a beep when the drop begins, a collision sound whenever the ball collides with an obstacle or the walls, and a softer collision sound when the ball lands in the sand on the floor. After hearing these auditory cues, participants view the image of where the ball ended up, and they can combine the auditory and visual information to figure out what happened. In the 'sound + ball occluded' condition, participants again first hear the ball being dropped. However, this time the box is only partially revealed. The final location of the ball remains occluded but the obstacles can be seen.

Intuitive Physics Engine

We model mental simulation in Plinko using an intuitive physics engine (IPE; Battaglia et al., 2013; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). The IPE is a cognitive model of physical reasoning that captures mental physical simulation by running noisy simulations in a computer physics engine.[1] The model is given an object-oriented representation of the scene, and a noisy simulator that updates the state of that scene according to approximate physical laws. Noise in the simulation process allows the model to capture uncertainty in human physical reasoning (Smith & Vul, 2013).

Figure 2b illustrates the IPE in the prediction task. To predict where the ball will fall, the model runs multiple noisy simulations (indicated by the dashed lines). We assume that people are uncertain about exactly how the ball is dropped, and what happens when it collides with the obstacles or the walls. To capture this uncertainty, we add noise to the simulations at two junctures (see middle panel): the angle of the ball's initial velocity when it drops from the hole, and the ball's velocity when it separates from an obstacle after a

[1]We implement the IPE with the 2D physics engine Chipmunk (`https://chipmunk-physics.net`), accessed with the python library pymunk (`https://www.pymunk.org`).

collision (Smith & Vul, 2013). The noise of the angle drop is represented with a single parameter representing the standard deviation of the Gaussian distribution on the drop angle, σ_d. The collision noise is represented with a pair of parameters, μ and σ_c, representing the mean and standard deviation for the Gaussian distribution from which a multiplier is drawn that's applied to the magnitude of the ball's velocity ($\mu = 1$ and $\sigma_c = 0$ represent deterministic collisions). The densities on the floor of the Plinko box reflect the distributions of simulated outcomes. Repeated simulation allows us to capture structure in participant prediction distributions such as bi-modality and skew.

Experiment 1: Prediction

Our first experiment motivates the Plinko domain as a test-bed for studying mental simulation in physical reasoning. The IPE closely captures the distribution of participant predictions, underlining the importance of mental simulation in physical prediction.

Human Results

Figure 2a shows participant data for one of the obstacle configurations. Small points represent participant clicks, densities represents the overall distribution of participant responses for the given hole, and the large connected circles represent the ground truth trajectory. Across all trials, the average absolute distance between participant predictions and the ground truth is 61.72 pixels (bootstrapped 95% confidence interval: [61.13, 62.33]). The distributions of participant responses are clearly structured, while also reflecting uncertainty in where the ball will land. This uncertainty can manifest as bimodality as in the case of hole 1 where participants are unsure whether the ball will end up to the left or right of the triangle. Noticeably participants tend to underestimate how far the ball travels after colliding with an obstacle. This is clearest in hole 2. The ground truth location of the ball is to the right, while the bulk of participant predictions are more toward the middle of the box.

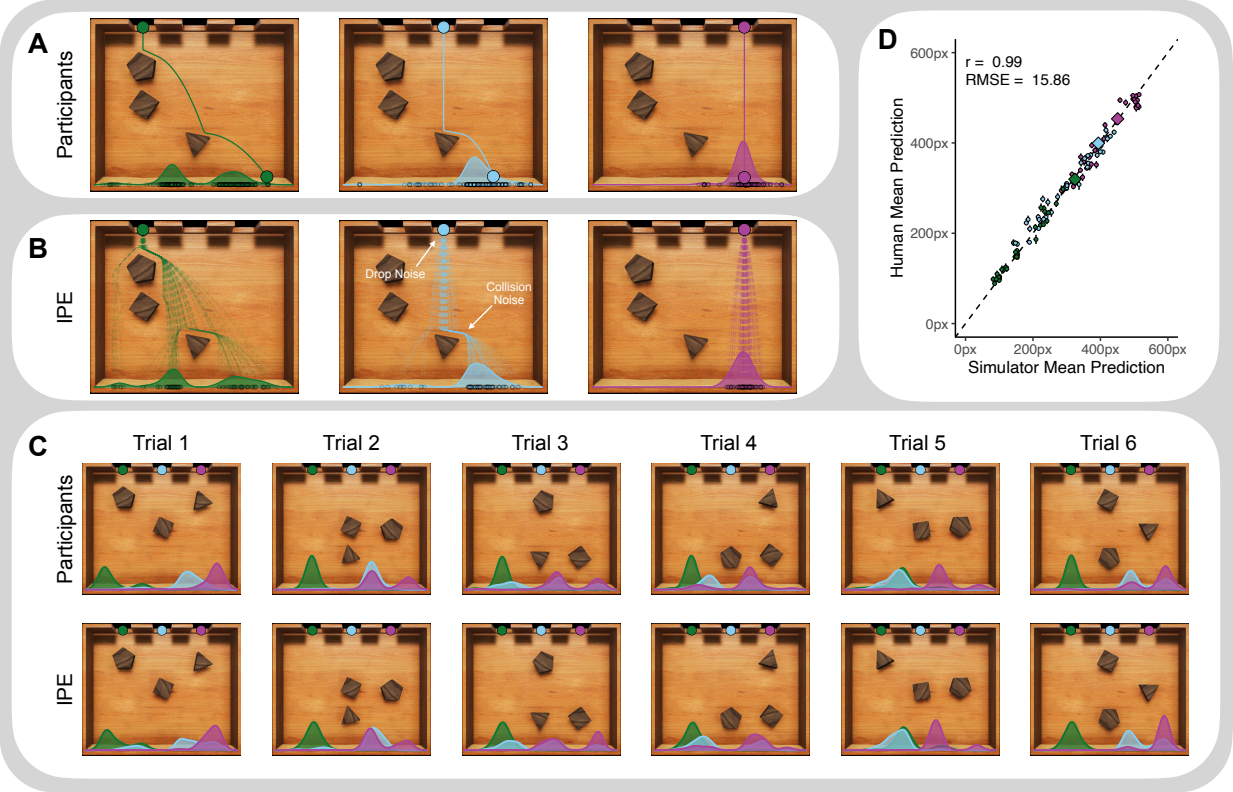

Figure 2

Evaluation of the IPE in the prediction task. ***A*** *Participant predictions together with kernel density estimates. The trajectories show where the ball actually ends up.* ***B*** *IPE predictions for the corresponding trials. Dotted paths represent simulation paths from the IPE.* ***C*** *Six additional sample trials showcasing participant and model predictions for all three holes. Overall the simulated distributions closely match the distributions of participant predictions across trials.* ***D*** *Scatterplot of model and participant mean predictions. Axes indicate pixel point of the mean prediction. The Plinko box is 600 pixels wide. The sample trial illustrated in* ***A*** *and* ***B*** *is shown as large diamonds. The IPE correlates strongly with participants' predictions.*

Modeling Physical Prediction

For each trial in the prediction task, the IPE runs 100 noisy simulations, producing a distribution of ball drop locations that can be compared to participant behavior (see Figure 2b). Figure 2c showcases six additional trial comparisons. The IPE consistently matches the fine-grained structure of the distribution of participant responses. Figure 2d summarizes the overall model performance, plotting the mean model prediction against mean participant prediction. Across trials, the IPE shows a high degree of correspondence

with participant data with $r = 0.99$ and RMSE = 15.86.

Mean prediction gives us a sense of whether the model captures the general location of participant predictions, but we would also like to assess the extent to which it captures the overall response distribution. To do so, we compute the Earth Mover's Distance (EMD, Rubner, Tomasi, & Guibas, 2000) between the model distribution and participant distribution. This metric is sensitive to additional structure in the distributions such as bimodality and skew, but it requires comparison with other models to evaluate goodness of fit. We compare the IPE with four lesioned simulation models and alternative models that don't use physical simulation.

No Drop Noise This model removes the drop noise variance from the IPE (σ_d). When simulating predictions in this model, the ball drops straight down every time.

No Collision Noise This model removes the collision noise parameters (μ and σ_c). Collisions are simulated deterministically.

No Noise Model This model removes all sources of physical noise from the IPE and runs a single deterministic simulation, reflecting the ground truth drop trajectory. We then fit a Gaussian distribution on the final location to capture variation in participant responses.

Under Hole Model This simple visual baseline predicts that human clicks will cluster normally directly under the location of the drop hole.

Statistical Models. We consider a set of statistical models to assess how much structure in the distribution of participant responses can be explained without explicit representation of an underlying simulation process. A powerful pattern associator could itself learn to simulate through multiple steps of non-linear transformation. Thus, we consider a sequence of models of increasing complexity, successively reducing the restrictions on the flexibility of the statistical transformation. Each statistical model transforms an input representation into the parameter space of a flexible mixture distribution, limiting the dimensionality of the prediction output while maintaining the flexibility to capture interesting structure in participant responses (e.g. multimodality). All

Table 1

Prediction Experiment Cross-Validation. Comparison of IPE performance on the prediction task against a series of lesions, statistical models, and heuristic alternatives. Model performance is summarized with the median performance across splits for correlation coefficient between prediction means (r), root mean square error between prediction means (RMSE), and earth mover's distance between overall human and model distributions (EMD). Across all metrics the IPE is the top performing model, followed by the 'no drop noise' lesion and the 'multi-layer non-linear' model.

Model	r	Δr	RMSE	ΔRMSE	EMD	ΔEMD
IPE	**0.99 [0.99, 0.99]**	–	**17.29 [14.53, 20.39]**	–	**20.62 [18.43, 23.60]**	–
No Drop Noise	0.98 [0.98, 0.99]	0.01 [0.00, 0.02]	27.34 [22.73, 32.70]	9.89 [5.22, 15.10]	30.03 [28.08, 32.35]	9.36 [6.79, 12.10]
No Collision Noise	0.93 [0.89, 0.95]	0.06 [0.04, 0.10]	49.35 [39.28, 63.18]	32.19 [23.16, 44.48]	47.85 [43.20, 51.60]	27.13 [23.16 31.06]
Deterministic Physics	0.89 [0.85, 0.93]	0.10 [0.06, 0.14]	66.21 [53.21, 78.82]	49.40 [37.21, 61.36]	55.32 [51.08, 61.20]	34.73 [30.80, 39.50]
Under Hole	0.84 [0.82, 0.87]	0.15 [0.12, 0.18]	69.24 [63.20, 73.81]	52.11 [45.03, 57.78]	67.19 [62.91, 70.14]	46.39 [41.41, 50.46]
Linear	0.70 [0.57, 0.78]	0.29 [0.21, 0.42]	96.91 [82.24, 118.09]	79.20 [65.04, 101.20]	87.33 [75.61, 101.69]	66.50 [54.81, 82.80]
Restricted Non-Linear	0.79 [0.72, 0.84]	0.20 [0.15, 0.27]	79.52 [69.62, 90.49]	62.06 [51.35, 73.46]	71.59 [65.02, 80.43]	51.00 [44.05, 60.35]
Multi-Layer Non-Linear	0.97 [0.95, 0.98]	0.02 [0.01, 0.04]	32.24 [25.12, 41.39]	14.66 [7.96, 24.11]	30.23 [24.81, 38.71]	9.42 [3.26, 18.56]

statistical models are pre-trained on a large set of simulated deterministic samples in the Plinko domain, and then fine-tuned on participant data. Details are included in Appendix A.

Linear Model This model computes a linear transformation from trial feature representation into the parameter space of the mixture model.

Non-Linear Restricted Depth This model computes a two-layer non-linear transformation – one linear layer followed by a GELU (Hendrycks & Gimpel, 2016) non-linearity followed by a linear transformation – from the input feature representation to mixture parameters. Adding non-linearity to the model allows it to learn more flexible transformations, but restricting its depth prevents the model from learning a multi-step "simulation-like" strategy to solve the problem.

Multi-Layer Non-Linear This model computes a multi-layer non-linear transformation from the image space to the mixture parameters. The model employs a resnet-like convolutional architecture (He, Zhang, Ren, & Sun, 2016; LeCun et al., 1989; Li, Liu, Yang, Peng, & Zhou, 2021) to better capture the spatial structure of the visual stimulus.

Prediction Evaluation

To evaluate the IPE against alternative models, we perform 100 split-half cross-validation runs summarized in Table 1. The IPE is the top-performing model for all three metrics, followed by the 'no drop noise' model and the 'multi-layer non-linear' model which perform similarly to one another. While the 'multi-layer non-linear' model does well, the other two statistical models perform even worse than the 'under hole' heuristic and the 'deterministic physics' model. Additionally, the successive improvement of the IPE over each of the model lesions suggests that both sources of noise (drop and collision) capture meaningful variation in participant behavior. This model comparison suggests that mental simulation is a plausible cognitive mechanism explaining participants' physical predictions in Plinko. The IPE provides the best explanation of participant behavior, and the closest alternatives either use simulation in more restricted forms ('no drop noise') or are mechanistically opaque ('multi-layer non-linear').

Experiment 2: Inference

Experiment 1 motivates mental simulation as a plausible cognitive mechanism for making predictions about the future in Plinko. In Experiment 2 we explore how people deploy this capacity to make inferences about what happened in the past. In physical environments, people often rely on both visual and auditory cues to figure out what happened. In this experiment, we demonstrate how the IPE coordinates these different sensory modalities via an underlying generative model. We compare participant behavior across three different sensory conditions (see Figure 1b): a 'no sound + ball visible' condition where participants only see the final location of the ball after a drop, a 'sound + ball visible' condition where participants first hear the sounds of the ball being dropped and then see the final location, and an 'sound + ball occluded' condition where participants hear the sounds of the ball drop and then see a partially occluded Plinko box revealing the obstacle locations but not the ball location. After being provided this condition specific evidence, participants indicate which hole they think the ball fell from.

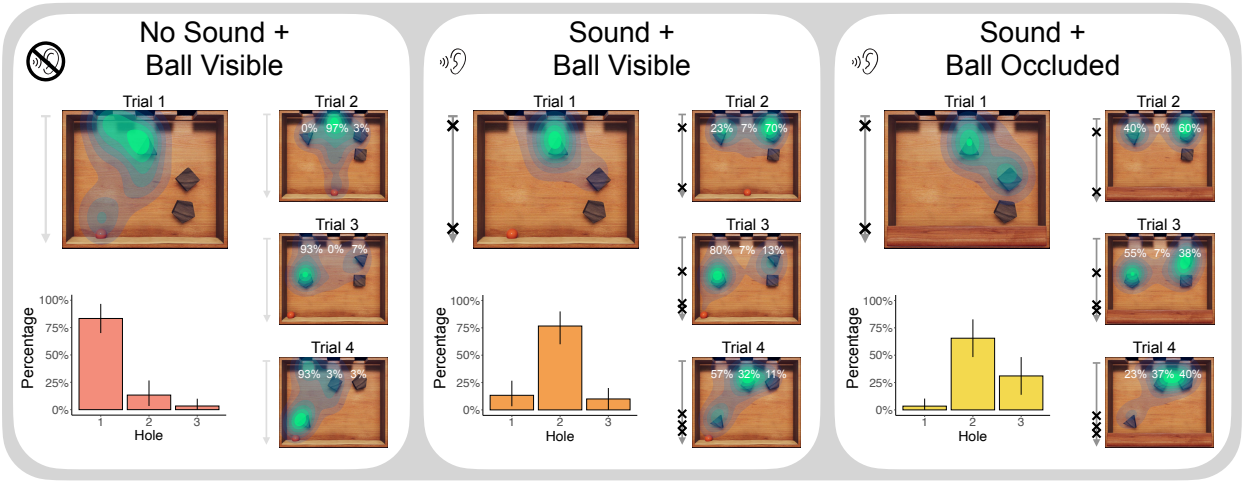

Figure 3

Participant behavior in the Inference task in four sample trials across all conditions. Gaze pattern is indicated by the heatmaps. For trial 1, judgments are shown as bar graphs. For the rest of the trials, judgment percentages are shown beneath the holes. Across the three conditions, different sources of evidence dramatically shift patterns of behavior, as participants examine different possibilities and make different selections in response to the different evidence profiles.

In addition to these inference judgments, we also track participants' eye-movements to better assess how they perform these inferences. Eye-data provides direct evidence for the role of mental simulation in the inference process, showcasing the particular hypotheses that participants entertain as they make their judgment. Examining differences in the distribution of eye-movement across conditions illustrates how different combinations of sensory evidence shape what hypotheses participants consider.

Human Results

Figure 3 illustrates participant behavior in four sample trials. In Trial 1, the ball was dropped from the middle hole and collided with the triangle obstacle at the top before it landed in the sand underneath the left hole. However participants do not view the drop itself. Instead they receive different combinations of auditory and visual cues depending on the condition. The different panels of Figure 3 illustrate how participants' behavior shifts with these different types of evidence. In the 'no sound + ball visible' condition, participants split their visual attention between two plausible hypotheses, hole 1 under

which the ball rests, and hole 2 where a collision with the triangle could also have given rise to the observed outcome. However, when it comes to their judgments, participants clearly favor the simpler hypothesis, the direct drop from hole 1. In the 'sound + ball visible' condition, this pattern changes. Because participants heard the ball collide with an obstacle, they now focus their attention on the triangle. Their judgments mirror this shift; the majority now favors hole 2 as the most plausible hypothesis. In the 'sound + ball occluded' condition the pattern shifts again. Here, they focus on the two hypotheses that could have resulted in the ball colliding with an obstacle. Compared to the other two conditions, the square under the right hole receives more attention, as participants can't see where the ball actually ended up. In this condition, participants' inferences are split between hole 2 and hole 3, though hole 2 is favored.

Trial 2 showcases how the presence or absence of auditory information can shape participants' behavior. In this trial, the ball drops from hole 3 and collides with the pentagon before landing directly beneath hole 2. In the 'no sound + ball visible' condition, participants focus their gaze at hole 2, and strongly favor this hole with their judgments. However, in the 'sound + ball visible' and 'sound + ball occluded' condition, participants gaze pattern shifts to examine the obstacles under hole 1 and 3. The auditory information reshapes the set of plausible hypotheses, and this is reflected in participant judgments.

Conversely, Trial 3 illustrates a situation where visual information crucially shapes participants' behavior. In this trial, the ball is dropped from hole 1 and collides with the pentagon before falling off to the left, colliding with the wall, and landing in the sand. Here, when participants can see the ball (in the 'no sound + ball visible' and 'sound + ball visible' conditions), they concentrate their gaze toward the left side of the Plinko box. It seems very unlikely that a drop from another hole would result in the ball falling so far to the side, and this is reflected in participants' judgments. However, in the 'sound + ball occluded' condition, participants are far less certain. They split their attention between the two holes with obstacles beneath and their judgments are much more divided.

Trial 4 illustrates a challenging trial, where additional auditory information doesn't necessarily help participants infer the correct hole. Here the ball drops from hole 1 and collides with the triangle, the wall, and then the floor each in quick succession. In the 'no sound + ball visible' condition, participants focus in on the triangle and tend to correctly infer hole 1. However, in the 'sound + ball visible' condition, participants hear the three collision sounds and many seem to consider the possibility that one of these collisions corresponds to the square, as suggested by the strong gaze density on the shape. Though the majority still select hole 1, a substantial minority selects hole 2 unlike in the 'no sound + ball visible' condition. In the 'sound + ball occluded' condition participants are even more uncertain, with the plurality now favoring hole 3.

Overall, participants are adept at inferring what happened with different combinations of sensory evidence. Across the three conditions, participants infer the correct drop hole at well above chance level (33%), with the highest accuracy in the 'sound + ball visible' condition (78% [77, 79]) followed by the 'no sound + ball visible' condition (72% [71, 73]) and then the 'sound + ball occluded' condition (63% [62, 65]). Participants' accuracy in 'sound + ball occluded' condition is particularly notable. In this condition, each source of sensory evidence (location of the obstacles, sounds of the collisions) provides little information on its own. It is remarkable then that participants are able to combine these individually uninformative sources to perform well above chance level. We verified that this difference from chance was credible with Bayesian regression analysis. We fit a logistic regression with uninformative priors predicting the participant accuracy in the 'sound + ball occluded' condition with a fixed intercept term as well as random intercepts for participants and trials. Analysis of the posterior reveals that the intercept is credibly above chance level (median: 72%, credible interval lower bound: 64%).

Modeling Physical Inference

We model physical inference in Plinko as Bayesian reasoning over an intuitive physics engine (IPE). Having observed some combination of evidence – the location of the

ball, the sound of the collisions, or both – a participant needs to determine the probability that the ball dropped from a particular hole. The posterior probability of a hole, h, given the ball location, b, and collision sounds, s, is proportional to the likelihood of that evidence given the hole, multiplied by the prior probability of the hole:

$$P(h|b,s) \propto P(b,s|h)P(h), \tag{1}$$

We assume the probability of the ball location is conditionally independent of the probability of the number of collision sounds given the hole, so we can further break down the likelihood:[2]

$$P(h|b,s) \propto P(b|h)P(s|h)P(h) \tag{2}$$

We assume the prior probability on the three holes is uniform. Thus, the key question to determine which hole has the highest posterior probability is how to construct the likelihood. In the previous experiment, we demonstrated that participants use their intuitive physical understanding to predict where the ball would fall if dropped from a given hole. The multiple predictions they provide can be thought of as samples from their internal simulation model. As demonstrated above, we can also sample repeatedly from the IPE to create probability distributions of the final ball location. This same model also creates a probability distribution for the auditory evidence. Running a simulation in the IPE will produce a particular sound pattern associated with the drop from a given hole. We represent the sound pattern as the number of collisions that can be heard when the ball is dropped. Running multiple simulations yields a likelihood of hearing a certain number of collisions given that the ball was dropped from a particular hole.[3]

[2]This assumption is not necessarily valid. Given a particular hole, knowing the number of sounds could provide information about the ball location and vice versa. However, we found that modeling the two sources of evidence jointly or independently did not meaningfully impact model performance, and assuming independence simplifies computation.

[3]Other features of the auditory signal, such as the timing of the collisions, could in principle be incorporated into the model's representation of the auditory signal. We found that considering the number of collisions worked best.

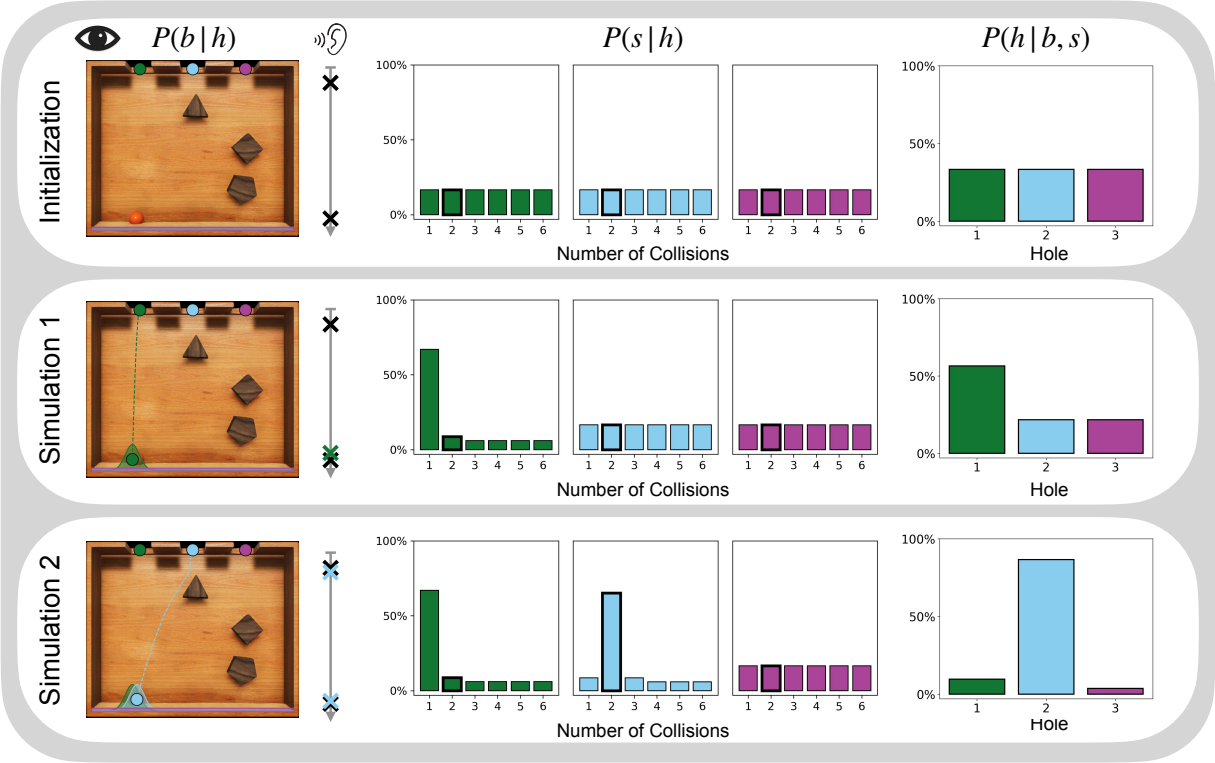

Figure 4

Illustration of multimodal integration with the Sequential Sampler. In this trial the ball drops from hole 2, colliding with the triangle before falling underneath hole 1. At the start the model doesn't favor any particular hole. It begins by simulating from hole 1. The location of the simulated outcome matches the visual evidence, but the simulated sounds do not match the auditory evidence. There is only one simulated collision when in reality there were two. The model then simulates from hole 2. The outcome of this simulation matches both the location of the ball and the number of sounds, leading to a strong swing in the posterior in favor of this hypothesis.

We represent these distributions with histograms updated through repeated simulations from the IPE. For the visual evidence, the IPE generates continuous outcomes along the floor of the Plinko box. We discretize these values to one of 600 possible outcome locations, the number of pixels along the floor of the Plinko box. Human mental simulation is not pixel precise. So we assume that a simulation that lands at a particular pixel location also increases the probability of neighboring pixels, and this increase drops off with increasing distance. Smoothing in this way reduces the number of samples necessary to construct a good estimate of the likelihood to a practical level (Cranmer, Brehmer, &

Louppe, 2020). We apply these smoothed updates with Gaussian kernels centered on the simulated outcome location. The bandwidth of the Gaussian kernel, σ_b, is a free parameter.

Similarly for the auditory evidence we maintain a distribution over the possible number of collisions.[4] If our model runs a simulation and there are two collision sounds, it will increment the count for this outcome. Similarly to the visual domain, we apply a Gaussian kernel to the update, capturing the idea that audition may be uncertain such that hearing a particular outcome also increases the probability of its neighbors. The width of this kernel is represented by another free parameter, σ_s.

Figure 4 illustrates how the IPE integrates evidence across the two sensory modalities. On this trial, the ball dropped from hole 2. In the conditions where participants have auditory evidence, they hear two collision sounds (one with the triangle early on, and one with the ground at the end). The visual and auditory likelihoods are initialized uniformly over the space of possible outcomes, so at the start, the model doesn't have a preference for any particular hole. In this example, the model begins by simulating from hole 1. The outcome of this simulation matches the visual evidence, but it contrasts with the auditory evidence, leading to a modest increase in the posterior probability of hole 1. In the next simulation, the model simulates from hole 2. The outcome of this simulation matches both the observed visual evidence as well as the auditory evidence, leading to a decisive swing in the posterior in favor of this hypothesis.

An important difference between the prediction and inference tasks in the Plinko domain is that, in the inference task, participants must consider simulations from multiple different possibilities. There is a limit to the amount of time a person will be willing to spend thinking about a single trial, and how participants allocate their simulations in that time could impact their judgment. The distributions of eye-movement in Figure 3 suggest that participants spend most of their time examining plausible hypotheses that could have

[4]We ran a large number of simulations from all holes on all trials in our stimulus set, and found the maximum number of collisions along a ball's trajectory from the hole to the ground was six.

given rise to the (condition-specific) evidence they received.

To capture this intuition, we develop a sequential sampling model that focuses simulation resources on plausible hypotheses. The Sequential Sampler runs a fixed number of simulations, selecting holes to consider in proportion to their posterior probability at a given point in time. With each simulation, the Sequential Sampler gains additional information about where the ball would have landed if it had been dropped from a particular hole and what pattern of sounds it would have made. Based on each simulated outcome, it updates the posterior probability (as illustrated in Figure 4), and decides what simulation to run next in proportion to its updated posterior probability.

Evaluating Model Inference

We evaluate the Sequential Sampler against participant judgments and eye-data. To assess whether this focused simulation strategy helps explain the pattern of participant behavior, we compare the Sequential Sampler to a Uniform Sampler that samples equally across the three holes. Both strategies produce judgment distributions that can be directly compared to the distribution of participant selections.

To evaluate our models against participant eye-data, we generate predictions of participant gaze distributions from our model behavior. Figure 5a illustrates our approach. We aggregate all participants' eye-data samples on a given trial to produce a heatmap representing the distribution of participant gaze. Then we construct a set of feature maps representing model behavior and salient visual features of the scene. We consider eight features in total. Four are visual features: the obstacle locations, the ball location, the hole locations, and the center of the screen.[5] The other four are dynamic features extracted from simulated model behavior: drop locations, obstacle collision locations, wall collision locations, and ground collision locations. Figure 5a showcases these feature maps. Notably, the dynamic features for the Sequential Sampler and the dynamic features for the Uniform Sampler differ for features like the drop and the obstacle collisions. The feature maps for

[5]Participants fixated on the center of the screen at the beginning of each trial.

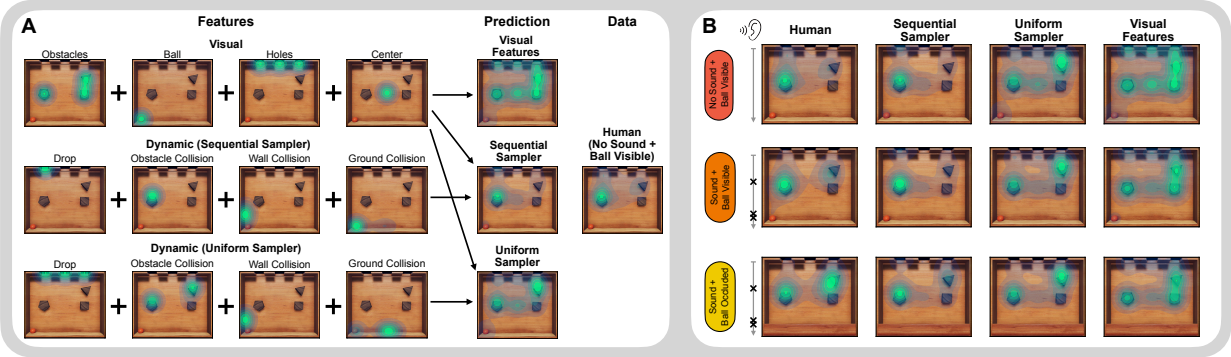

Figure 5

A *Visual illustration of the regression method for predicting the distribution of participant eye-data from model simulation behavior. The heatmap on the right represents the distribution of participant gaze for the trial, computed by aggregating all participant eye data on the given trial. We predict this distribution by constructing a set of feature maps representing various visual features of the trial (top row), as well as the dynamic simulated behavior of the model (middle and bottom row). We combine the visual and the model-specific dynamic features using linear regression to generate predictions for our two simulation models. We also compute a visual features heuristic using just the visual features.* ***B*** *Human heatmaps and model predictions for the same trial in all three conditions. In the 'no sound + ball visible' condition and the 'sound + ball visible' condition, participants and the Sequential Sampler focus their attention on the left of the Plinko box where the ball has landed off to the side. The Uniform Sampler and the Visual Features model are more diffuse in their allocation of attention. In the 'sound + ball occluded' condition, participants' gaze is not anchored to the left by the presence of the ball. The Uniform Sampler is better able to capture this pattern, though the Sequential Sampler still performs comparably.*

the Sequential Sampler concentrate their mass toward hole 1 and the resulting collisions, while the feature maps for the Uniform Sampler distribute mass evenly across the different holes and their simulation paths.

With these feature maps we generate predictions of the participant distribution with linear regression (for formal specification, see Data Analysis in Methods). We generate predictions for the Sequential Sampler and Uniform Sampler by combining the visual features with the (model specific) dynamic features. We also generate predictions for a visual features heuristic that only uses the visual features. We fit separate regressions for the Sequential Sampler and Uniform Sampler in each condition. Figure 5b illustrates the human heatmaps alongside the model predictions for a single trial in each of our three

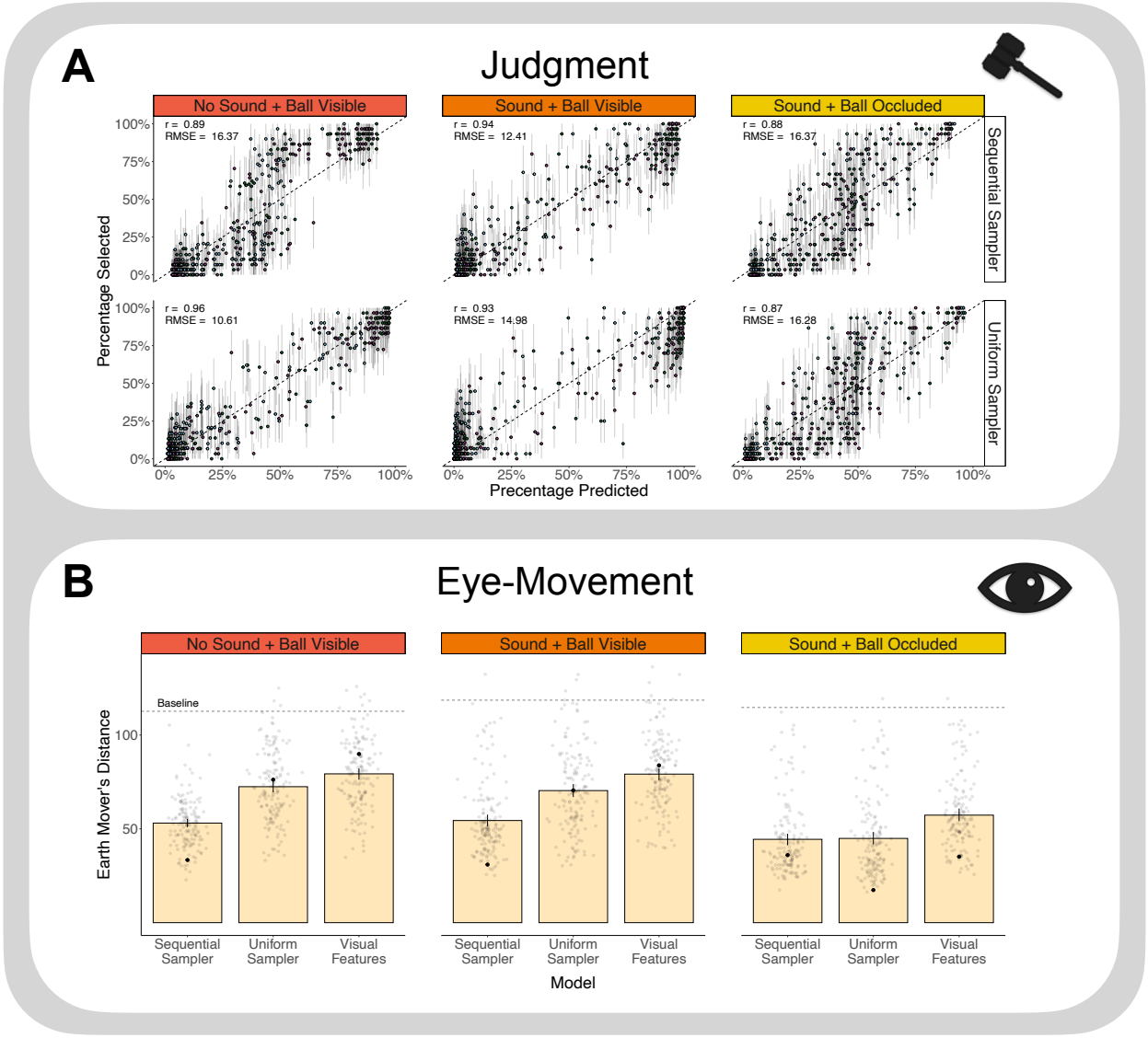

Figure 6

Model performance of the Sequential Sampler and Uniform Sampler across the three experimental conditions. ***A*** *Judgment data. The two sampling strategies perform similarly in explaining participants' hole choices.* ***B*** *Eye-movement data. The average Earth Mover's Distance (lower is better) between the participant distributions and the model prediction is substantially lower for the Sequential Sampler in the 'no sound + ball visible' and 'sound + ball visible' conditions. All three models perform similarly in the 'sound + ball occluded' condition. The dotted line represents a baseline comparison computed by evaluating the EMD between each participant distribution and a uniform density on the Plinko box and calculating the average. Darkened point represents the sample trial illustrated in Figure 5b.*

conditions. We evaluate how well the models predict participant distributions using the Earth Mover's Distance.

Figure 6a summarizes overall model performance on the judgment data. Both sampling strategies do well at predicting participant judgments. To meaningfully distinguish the models, we must consider the eye-data. Figure 6b shows model performance on this data. In the 'no sound + ball visible' condition and the 'sound + ball visible' condition, the Sequential Sampler outperforms the Uniform Sampler and the Visual Features heuristic, while in the 'sound + ball occluded' condition, the models perform more equivalently with a slight edge for the simulation models over the visual features model.

Figure 5b illustrates why the Sequential Sampler has an advantage over the alternatives in the first two conditions, and why performance is more even in the 'sound + ball occluded' condition. In this trial, the ball is dropped from the hole 1 and falls to the left side of the box far away from the other holes. In the conditions with sound, participants hear three collisions (one with the pentagon, one with the wall, and one with the ground). In the 'no sound + ball visible' and 'sound + ball visible' conditions, participants' gaze is drawn toward the side of the screen where the ball is located, and they focus most heavily on the obstacle along the simulation path (the pentagon). In the 'sound + ball occluded' condition, participants hear the three collisions but don't see the location of the ball, so they spend more time deliberating among the two trajectories that contain obstacles (hole 1 and hole 3). In the 'no sound + ball visible' condition and the 'sound + ball visible' condition, the Sequential Sampler better captures this emphasis on the most plausible hypothesis (see black point in Figure 6b). In contrast, the Uniform Sampler and the Visual Features Model spread their attention more broadly across the different objects. In the 'sound + ball occluded' condition, participants spread their attention more uniformly. The three models perform comparably here.

We evaluated model performance with 100 split-half cross-validation runs. In addition to the Sequential Sampler and Uniform Sampler, we considered two heuristic models, one specific to the judgment data, and the other to the eye-data. For the judgment data, we implemented a closest-hole prediction model. For a given inference trial, this

Table 2

Summary of the inference results, overall and cross-validation. We compare the Sequential Sampler against the Uniform Sampler, a 'closest hole' heuristic, and a 'visual features' heuristic. For each condition, we specify model performance on the judgments with correlation (r) and root mean square error (RMSE), and on the eye-data with Earth Movers Distance (EMD). Cross-validation summarizes performance on the test set of 100-split halves with median performance, 2.5, and 97.5 percentiles.

	Vision Only			Vision and Sound			Occlusion and Sound		
Overall Results									
Model	Judgment r	Judgment RMSE	Eye-Movement EMD	Judgment r	Judgment RMSE	Eye-Movement EMD	Judgment r	Judgment RMSE	Eye-Movement EMD
Sequential	0.89	16.37	53.07	0.94	12.41	54.44	0.88	16.37	44.42
Uniform	0.96	10.61	72.46	0.93	14.98	70.39	0.87	16.28	44.90
Closest Hole	0.72	24.49	–	0.57	29.76	–	–	–	–
Visual Features	–	–	79.31	–	–	79.17	–	–	57.30
Cross-Validation									
Sequential	0.89 [0.87, 0.91]	16.58 [15.01, 18.42]	52.94 [50.55, 55.09]	0.94 [0.92, 0.96]	12.79 [10.78, 14.71]	54.06 [51.09, 57.21]	0.87 [0.84, 0.90]	16.89 [15.06, 19.87]	44.47 [41.16, 48.04]
Uniform	0.96 [0.90, 0.97]	10.63 [9.43, 15.85]	72.10 [69.56, 74.21]	0.93 [0.91, 0.95]	14.84 [12.93, 16.50]	69.46 [66.29, 73.04]	0.87 [0.85, 0.90]	16.46 [14.89, 17.66]	45.25 [42.16, 49.13]
Closest Hole	0.72 [0.66, 0.79]	24.55 [23.26, 26.35]	–	0.57 [0.49, 0.66]	29.95 [27.15, 32.39]	–	–	–	–
Visual Features	–	–	78.73 [76.43, 81.22]	–	–	77.93 [75.19, 81.20]	–	–	57.14 [54.29, 62.23]
Differences									
Comparison	Δr	ΔRMSE	ΔEMD	Δr	ΔRMSE	ΔEMD	Δr	ΔRMSE	ΔEMD
Sequential − Uniform	−0.06 [−0.09, −0.01]	5.98 [1.08, 7.71]	−18.92 [−21.46 −16.43]	0.00 [−0.02, 0.02]	−2.26 [−3.77, 0.60]	−15.13 [−18.11, −12.54]	0.00 [−0.02, 0.01]	0.30 [−0.25, 2.77]	−0.89 [−2.71, 0.88]
Sequential − Closest Hole	0.17 [0.11, 0.22]	−7.88 [−9.75, −6.17]	–	0.36 [0.29, 0.45]	−17.16 [−19.62, −14.72]	–	–	–	–
Sequential − Visual Features	–	–	−25.76 [−28.39, −23.20]	–	–	−23.77 [−26.15, −21.30]	–	–	−13.02, [−15.19, −10.95]

model predicts the hole that is closest to the observed location of the ball. We fit a softmax function on these closest hole predictions in order to maximize the likelihood of participant selections under this model. Because this heuristic relies on observation of the location of the ball, it doesn't make predictions in the 'sound + ball occluded' condition.

For the eye-movement data, we use the visual features heuristic presented above. We fit the visual features regression on the train trials and then generate predictions for the test. We evaluate the model using the Earth Movers Distance between test predictions and participant eye-movement distributions.

Table 2 shows the overall and cross-validation results. The comparison between the Sequential Sampler and the Uniform Sampler in cross-validation largely reproduces the pattern we see with the overall results. The two models perform similarly when it comes to judgments, but for the eye-movement data, the Sequential Sampler outperforms the Uniform Sampler in the 'no sound + ball visible' and 'sound + ball visible' conditions. The two models perform more similarly in the 'sound + ball occluded' conditions. Compared to the heuristic models, the Sequential Sampler strongly outperforms the closest hole heuristic and the visual features heuristic.

Discussion

In order to navigate a complex environment, people must integrate multiple forms of sensory evidence in a common representation. Structured causal knowledge in the form of an intuitive theory provides an ideal grounding in which to coordinate these disparate kinds of sensory experience. Through mental simulation, an intuitive theory can apply general causal knowledge to specific circumstances, providing an explanation for how a particular pattern of multimodal evidence emerged in a given situation. This powerful cognitive capacity is a cornerstone of human intelligence, underlying our general ability to understand and explain the world we inhabit.

In this work, we developed a novel paradigm and modeling approach to examine this ability in the process of physical inference. We introduced the Plinko domain, a flexible intuitive physics setting where participants perform prediction and inference. In the prediction task, participants see the drop location of a ball in the Plinko box and must determine where it will land. We model participants' predictions in this task with an Intuitive Physics Engine (IPE, Battaglia et al., 2013). The IPE runs multiple simulations in a noisy simulator to generate a distribution of predictions for each hole in the Plinko box. The distribution of model predictions closely matches the participant response distribution, outperforming several alternative models. This finding provides evidence that mental simulation is critical for prediction in Plinko and builds on a growing literature demonstrating the success of the noisy simulator approach in modeling human physical judgments (Bates et al., 2019; Gerstenberg et al., 2021; Smith et al., 2013; Ullman et al., 2017; K. W. Wong et al., 2023; Zhou et al., in press).

In the inference task, we examine how people use their intuitive simulator to coordinate evidence from different sensory modalities. In three different experimental conditions, we present participants with different forms of sensory evidence, and they must determine which hole the ball fell from. In the 'no sound + ball visible' condition, participants see where the ball landed and the location of the obstacles. In the 'sound +

ball visible' condition, participants first hear the sounds of the ball drop while the Plinko box is covered, and then see the uncovered Plinko box with the ball on the ground. Finally, in the 'sound + ball occluded' condition, participants again hear the sounds of the ball drop, and then view a partially revealed Plinko box that reveals the locations of the obstacles but masks the final location of the ball. In addition to judgments, we also track participant eye-movement while they perform the task. Gaze patterns reflect the simulated physical trajectories that participants consider, providing insight into the underlying cognitive process guiding their judgment.

To model participant behavior in the inference setting, we extend the simulation approach from the prediction task. We simulate with the IPE to construct a likelihood distribution representing the probability of observing particular patterns of visual and auditory evidence given that the ball was dropped from a particular hole. We then use Bayesian inference to compute the posterior probability of each hole assuming a uniform prior. We consider two different simulation strategies with the IPE, a uniform sampler which simulates from each hole equally, and a sequential sampler which simulates from each hole in proportion to its posterior probability. While the two strategies perform comparably at explaining participant judgments, the sequential sampling strategy better captures the distribution of participant eye-movement, reflecting participants' tendency to focus toward hypotheses of interest.

Our work here connects with a larger literature examining how humans integrate multisensory experience in a unified representation of their environment (Alais & Burr, 2019; Stein & Meredith, 1993). Here, we build on a growing trend in the perception literature highlighting the importance of latent causal structure in guiding multisensory integration (Körding et al., 2007; Rohe & Noppeney, 2015; Shams & Beierholm, 2010). Causal structure in Plinko is more complex than traditional psychophysical tasks, but humans are nonetheless quite adept at inference in this domain. By relating sensory cues to a latent, simulated causal process, humans can perform impressive feats of multisensory

inference, beyond what would be possible with each modality independently. This is most evident in the 'sound + ball occluded' condition, where participants must combine partial visual cues with auditory information. It is hard to see how a traditional cue combination approach (Alais & Burr, 2004; Ernst & Banks, 2002), where sensory signals are combined in a linear additive fashion, could explain this behavior. In this condition, each signal provides little to no information individually about the drop location. The location of the obstacles are uninformative without the sounds, and likewise the sounds on their own give very little information without some sense of where the obstacles are located. Only when the two signals are combined in a structured causal model representing the physical dynamics of the world, is it clear how this information constrains the drop location. Causal knowledge allows sensory signals to provide information beyond their simple sum, helping to explain how human thinkers navigate the intricate complexity of their environment.

Limitations and Future Directions

Our work here opens several avenues for further investigation. We examined participant behavior at a fairly coarse level, aggregating trial data across participants and, in the case of eye-movement, across time as well. Unpacking this aggregation to reveal individual eye trajectories, as well as how those gaze patterns correspond with particular judgments, has the potential to reveal additional information about the underlying cognitive process. In Plinko, cognition and perception are closely intertwined: perceptual information gathering is performed in service of the higher level goal of refining simulation uncertainty. Combining our simulation approach with models of visual search behavior (Butko & Movellan, 2010; Najemnik & Geisler, 2005; Vasilyev, 2018) could deepen our understanding of why participants look where they look. Examining individual visual trajectories could also explain why two participants arrive at different judgments on the same trial.

Digging more deeply into individual differences raises additional questions about how participants vary in this task. One notable axis of difference is the amount of time

participants take to come to a decision. Supplemental response time data in the inference task reveals substantial individual variation. Both our Sequential Sampler and our Uniform Sampler perform a fixed number of simulations on all trials. But it is plausible that different participants dedicate different levels of simulation resources in the task, or that the same participant exerts more or less effort between trials. Unpacking this decision process in greater detail, understanding how participants decide when they have seen enough, as well as what utilities participants consider (e.g. hole probability, expected information gain, etc.) when deciding which hypothesis to simulate is an important direction for future work. In this regard, Evidence Accumulation Models (EAMs, Evans & Wagenmakers, 2020; Ratcliff, Smith, Brown, & McKoon, 2016) provide a promising source of inspiration. Combining these tools with our own simulation models could help us better describe how cognitive effort features into people's decisions and how that shapes individual responses.

We model mental physical simulation in this work with the IPE. The model does a good job of predicting simulation behavior in our task, and has also seen success at modeling intuitive physical judgments in several other domains. However, critical perspectives on the IPE have highlighted certain aspects of the model that seem unlike human physical reasoning (Ludwin-Peery, Bramley, Davis, & Gureckis, 2021). Unlike computational physics engines, which maintain detailed representations of the entire simulated domain, humans seem to make do with partial simulations focused on key objects and events of interest (Balaban & Ullman, 2025; Bass, Smith, Bonawitz, & Ullman, 2021; Bigelow, McCoy, & Ullman, 2023; Hegarty, 2004). This partiality is reflected in participant eye-movements. In Plinko participants focus their visual attention to key moments of the simulation (primarily object collisions). However, the IPE always rolls out complete simulations from start to finish. Unpacking this relationship in greater detail is important for more faithfully modeling human physical reasoning, and also clarifying the relationship between simulation and eye-movement which we explore in this work.

Reconsideration of the simulator raises broader questions about how people perform

multimodal inference in other causal domains. Plinko is particularly well suited for examining this capacity in physical reasoning, but how do people perform inferences and combine evidence in other causal systems? And how do different types of evidence support those inferences in different domains? In social settings, people are adept at inferring folk psychological stories explaining why a person acted in one way or another (Baker, Saxe, & Tenenbaum, 2009; Gerstenberg & Tenenbaum, 2017; Jara-Ettinger et al., 2016; Wu, Sridhar, & Gerstenberg, 2022). In addition to primary sensory experience, language is a crucial source of information constraining the social stories we imagine (L. Wong et al., 2023). Developing models and methods that allow us to examine these processes is critical for more fully capturing how people understand the causal world.

Conclusion

In this work, we investigated how people perform multimodal inference in intuitive physical reasoning using mental simulation. We introduce the Plinko paradigm, a physical reasoning task where participants perform predictions and inferences. We demonstrate that a model that runs simulations in an Intuitive Physics Engine (IPE) does an excellent job of matching participant predictions. We go on to illustrate how a simulation model can capture participant judgments and eye-movement in an inference task. In three conditions involving different combinations of visual and auditory evidence, our model matches participant behavior in both data signals. By coordinating sensory evidence in an underlying causal representation of the domain, the model is able to go beyond simple combinations of sensory experience to explain complex higher level inferences. This ability to reconstruct the past from partial, multimodal evidence in the present is foundational to human intelligence. Our work here takes us one step closer to understanding this fundamental cognitive capacity.

Methods

Prediction Task

This experiment was approved by MIT's IRB (#0812003014). We recruited 45 participants (*age*: $M = 37$, $SD = 11$, *gender*: 21 female, 24 male) on Amazon Mechanical Turk using Psiturk (Gureckis et al., 2016). Participants first consented to participate and were then introduced to the Plinko box. Participants received instructions that their task would be to predict where the ball would land. They viewed four videos of the ball being dropped into the Plinko box, answered a set of comprehension check questions, and then proceeded to the main task.

Participants were presented with 122 trials like the one illustrated in Figure 1A. The first two trials were training trials, the remaining 120 trials were test trials and were presented in randomized order. Participants clicked on the screen 10 times indicating where they thought the ball would land. Upon each click, a red semi-transparent dot appeared at the horizontal location of the click on the floor of the Plinko box. Participants were told they could click in the same location multiple times to indicate their confidence. Participants viewed 40 different Plinko box configurations and made predictions for each of the three holes in each configuration. Due to trial randomization, trials for different holes in the same configuration did not in general occur in sequence.

Fitting the IPE

We fit the parameters of the IPE using a grid search. We considered values ranging from 0 to 1 in increments of 0.1 for each of our three noise parameters, drop noise (σ_d), collision mean (μ), and collision standard deviation (σ_c). For each parameter setting, we generated model simulations for all three holes on all trials and computed kernel density estimates (KDE) from the simulation outcomes from each of these holes. We then evaluated the likelihood of participant predictions for those trials under those KDEs and summed the total log likelihood across all trials. We fit the KDE bandwidth for each setting of the IPE parameters to maximize the likelihood of participant clicks under the

corresponding KDEs (for a given parameter setting all the KDEs for all trials had the same bandwidth). We found an optimal value of 0.3 for σ_d, 0.0 for μ, and 0.7 for σ_c, and 36.2 for the KDE bandwidth.

Inference Task

This experiment was approved by Stanford's IRB (#48663). We recruited 90 participants (*age*: $M = 24, SD = 8$, *gender*: 51 female, 39 male, *race*: 38 Asian, 35 White, 4 Black/African-American, 2 Pacific Islander, 6 Multiracial, 5 other/unclear) through Stanford's student experiment participation portal. Participants were compensated for their participation with course credit. After consenting, participants received instructions on the task. We introduced them to the Plinko box and showed them 6 videos depicting drops from the holes. For participants in the 'sound + ball visible' and 'sound + ball occluded' conditions, the videos were accompanied by sounds of the drop, but in the 'no sound + ball visible' condition, participants only saw the drop. Participants were instructed that their task would be to infer where the ball had fallen from based on condition-dependent forms of sensory evidence. In the 'no sound + ball visible' condition, participants were only provided with the image depicting the endstate of the drop. In the 'sound + ball visible' condition, participants first heard the drop unfold while looking at the covered Plinko box, and then were provided with the visual stimulus as well. In the 'sound + ball occluded' condition, participants were again presented with the auditory cue, followed by a partially occluded version of the trial stimulus that revealed the locations of the trial obstacles but not the location of the ball (see Figure 1).

After receiving instructions, we calibrated the eye-tracker. Participants rested their chins on a headrest 51.2 centimeters from the center of the screen. We collected data using an Eyelink 1000 with a sampling rate of 1000 Hz. We calibrated using Eyelink's default 9-point calibration routine. Participants then had a chance for a brief break before proceeding to the experiment trials. Trial stimuli were 23.2 centimeters wide (12.8 degrees of visual angle) and 19.5 centimeters tall (10.8 degrees of visual angle). Participants

completed 2 practice trials before continuing to the main battery of 150 trials. The order of the main trials was randomized. Participants started each trial with drift correction, fixating in the center of the screen before proceeding to the trial stimulus. After hearing the auditory cues (if any) and upon viewing the visual stimulus, participants could complete the trial at any point by pressing the 1, 2, or 3 number key on the keyboard, corresponding to the different holes. Every 30 trials participants took a break to rest their eyes.

Data Preparation

We first cleaned data to remove outliers. We calculated mean and standard deviation of response time for each condition. Trials where the response time exceeded the mean by more than three standard deviations were excluded from analysis. We removed 76 trials from the 'no sound + ball visible' condition (4424 trials remaining), 84 from the 'sound + ball visible' condition (4416 remaining), and 19 from the 'sound + ball occluded' condition (4484 remaining).

Participants begin each trial with drift correction, where they are required to fixate at the center of the screen in order to proceed to the trial. For participants in the 'no sound + ball visible' condition, drift correction took place directly before the onset of the visual stimulus. However, for participants in the 'sound + ball visible' and 'sound + ball occluded' conditions, drift correction took place before the the beginning of the auditory stimulus, allowing participants to stray from center fixation before we began recording their eye-movement. In order to standardize the impact of center fixation across conditions, we examined how the average distance of eye-data samples from the center of the stimulus developed over the trial. In each condition, we calculated the distance of every eye-data sample from the center of the stimulus and calculated the average distance in 100 ms bins across all trials in the condition. Participants in 'no sound + ball visible' condition started on average quite close to the center screen ($\sim$10 pixels), while participants in the 'sound + ball visible' and 'sound + ball occluded' conditions started on average quite a bit further ($\sim$60 pixels in both). However between 300 and 400 ms, participants in all three conditions

coalesced to an average distance of around 100 pixels from the center. Thus, in order to standardize, we removed the first 300 ms of eye-movement data in all three conditions from analysis.

Fitting the Sequential Sampler

We used the noise parameters fitted in the prediction task for the IPE. To fit the remaining model parameters of the Sequential Sampler, we evaluate our model on participant judgments and eye-data in each condition. To assess model performance on the judgment data, we take the model posterior distribution computed for a given trial and compute the log likelihood of participant selections under that distribution. Judgment performance for a particular parameter setting in a given condition is calculated as the sum of the log likelihood of participant selections across all trials.

To evaluate model performance on the eye-movement data, we assess to what extent the behavior of the model can be used to predict the overall distribution of participant eye-movement. For each trial in a condition, we compute a two-dimensional kernel density estimate of participants eye-data, aggregating all eye-data samples across participants. The KDE has a fixed bandwidth ($bw = 50$ pixels) with a Gaussian kernel. We then create a series of feature maps representing visual features in the Plinko box and model behavior that we can use to predict the distribution of eye-movement (see Figure 5). The four visual features we consider are the obstacle locations, the final location of the ball, the hole locations, and the center of the Plinko box. The four dynamic features representing simulation model behavior are the locations of the simulated drops, the locations of the obstacle collisions, the locations of the wall collisions, and the location of the ground collisions. To create the feature maps representing each of these predictors, we compute kernel density estimates where the feature locations serve as the data points. For example, for the obstacle location feature map, we take the center points of each of the three obstacles and create a KDE where the distribution mass is clustered around those center points. For the obstacle collision feature map we do the same, but the distribution is now

centered on the locations of the simulated collisions. The obstacle feature map is dependent only on the visual features of the trial, but the obstacle collision feature map is dependent on the dynamic behavior of the simulator.

Once we have these feature maps, we fit a regression predicting the kernel density values of the eye-data KDE from the feature map KDEs. Each kernel density estimate implies a density value at each pixel point of the trial stimulus. We can formulate this like so:

$$y_{trc} = \beta_0 + \beta_1 x_{1trc} + ... + \beta_8 x_{8trc} + \epsilon_{trc} \tag{3}$$

Here, y_{trc} represents the eye-data kernel density value for a particular trial (t) at a given row (r) and column (c). x_{itrc} represents the corresponding feature map value for each of our eight features. We fit eight coefficients for each of these feature maps as well as an intercept and a noise value. For each regression, we have a total of 10 fitted parameters, where the predictor coefficients (β_1 through β_8) roughly correspond to the importance of the corresponding feature in predicting the overall distribution of participant eye-movement. See Figure 5a for a visual illustration of the regression method.

We use this fitted regression to compute model predictions of the eye-movement data. We take these predictions and transform them into valid distributions by setting any predicted values below zero to zero, and renormalizing the distribution so that it sums to 1. We can then compare these predicted distributions to the distribution of participant eye-data using the Earth Movers Distance (EMD Rubner et al., 2000). Computing EMD involves solving an optimal transport problem with linear programming. Computing the distance for distributions with large domains becomes intractable. Our participant KDEs and model predictions are 500 $\times$ 600 pixels. In order to compute the EMD, we downsample the activations, computing the average density value in 20 $\times$ 20 blocks. We then compute the EMD between these downsampled representations which are 25 $\times$ 30. Finally, to compute the model performance for a particular parameter setting in a given

condition, we calculate the average EMD across all trials in the condition.

We evaluate model behavior for every parameter setting (all combinations of σ_b and σ_s) on each data signal in each condition. We then rank model performance across parameter settings within each signal-condition combination, resulting in 6 different performance rankings. To determine the overall model performance for a particular parameter setting we compute the average rank. Following this procedure we found an optimal visual bandwidth (σ_b) of 100 pixels and an optimal auditory bandwidth (σ_s) of 0.4 collision sounds.

References

Ahuja, A., Rodriguez, N. Y., Ashok, A. K., Serre, T., Desrochers, T. M., & Sheinberg, D. L. (2024). Monkeys engage in visual simulation to solve complex problems. *Current Biology*, *34*(24), 5635–5645.

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

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, *14*(3), 257–262.

Alais, D., & Burr, D. (2019). Cue combination within a bayesian framework. In *Multisensory processes: The auditory perspective* (pp. 9–31). Springer.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

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

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413–424.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLOS Computational Biology*, *15*(7), e1007210.

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.

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

Beller, A., Xu, Y., Linderman, S., & Gerstenberg, T. (2022). Looking into the past:

Eye-tracking mental simulation in physical inference. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

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

Butko, N. J., & Movellan, J. R. (2010). Infomax control of eye movements. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 91–107.

Chen, Y.-C., & Scholl, B. J. (2016). The perception of history. *Psychological Science*, *27*(6), 923–930.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, *117*(48), 30055–30062.

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.

Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach. *PLoS computational biology*, *11*(11), e1004610.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Evans, N. J., & Wagenmakers, E.-J. (2020). Evidence accumulation models: Current limitations and future directions. *Quantitative Methods for Psychology*, *16*(2), 73-90.

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*(5), 936.

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. (2018). What happened? reconstructing the past from vision and sound. *Proceedings of the 40th Annual*

Conference of the Cognitive Science Society.

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

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? the amount of mental simulation tracks uncertainty in the outcome. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 866–871). Austin, TX: Cognitive Science Society.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision research*, *39*(21), 3621–3629.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(10), 785.

Kim, M., & Schachner, A. (2025). Sounds of hidden agents: The development of causal reasoning about musical sounds. *Developmental Science*, *28*(4), e70021.

Körding, K., Beierholm, U., Ma, W., Quartz, S., Tenenbaum, J., & Shams, L. (2007). Causal inference in multisensory perception. *PLOS ONE*, *2*(9), e943.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building

machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541-551.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, *33*(12), 6999–7019.

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

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, *20*(4), 260–281.

Rohe, T., & Noppeney, U. (2015). Cortical hierarchies perform bayesian causal inference in multisensory perception. *PLoS biology*, *13*(2), e1002073.

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*, 99–121.

Shams, L. (2012). Early integration and Bayesian causal inference in multisensory perception. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes.* CRC Press, Boca Raton, FL.

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in cognitive sciences*, *14*(9), 425–432.

Smith, K. A., Dechter, E., Tenenbaum, J., & Vul, E. (2013). Physical predictions over time. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 1342–1347).

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. MIT press.

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.

Vasilyev, A. (2018). Optimal control of eye movements during visual search. *IEEE Transactions on Cognitive and Developmental Systems*, *11*(4), 548–559.

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

Wong, K. W., Bi, W., Soltani, A. A., Yildirim, I., & Scholl, B. J. (2023). Seeing soft materials draped over objects: A case study of intuitive physics in perception, attention, and memory. *Psychological Science*, *34*(1), 111–119.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., & Tenenbaum, J. B. (2023). From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2008). Human trimodal perception follows optimal statistical inference. *Journal of vision*, *8*(3), 24–24.

Wu, S., Sridhar, S., & Gerstenberg, T. (2022). That was close! a counterfactual simulation model of causal judgments about decisions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Yildirim, I., & Jacobs, R. A. (2014). Learning multisensory representations for auditory-visual transfer of sequence category knowledge: a probabilistic language of thought approach. *Psychonomic Bulletin & Review*, *22*(3), 673–686.

Ying, L., Xu, D., Zhang, A., Collins, K. M., Siegel, M. H., & Tenenbaum, J. B. (2025).

What's in the box? reasoning about unseen objects from multimodal cues. *arXiv preprint arXiv:2506.14212*.

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (in press). Mental Jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*.

Appendix A

Statistical Model Architecture and Training

To model the distribution of human predictions across spatial decision-making tasks, we parameterized a *Beta Mixture Model* (BMM) using linear, MLP, and convolutional architectures. Each model receives task-specific input features and outputs the parameters of a BMM that describes the probabilistic distribution over human responses, where the human responses are binned into one of 600 discrete bins ranging over the possible drop locations in each data image.

Model Architectures

We considered three model architectures. Other than for the distribution parameter prediction, each dense weight matrix in the models was preceded by a Layer Normalization (Ba, Kiros, & Hinton, 2016) using PyTorch's default hyperparameters and followed by a Dropout layer (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) with 0.5 probability to drop. The final hidden layer projects to three separate linear output heads:

- **Mixture Weights (c)**: Softmax-normalized vector of length $K = 10$ over mixture components.
- **Alpha (α)**: Positive-valued shape parameters of the beta components, obtained by exponentiating the output of a fully connected layer.
- **Beta (β)**: Same as α, with separate learned weights.

The final mixture model is represented as a *MixtureSameFamily* distribution over beta components:

$$p(y \mid x) = \sum_{k=1}^{K} c_k(x) \cdot \text{Beta}(y \mid \alpha_k(x), \beta_k(x))$$

where $y \in (0, 1)$, c_k is the mixture weight, and α_k, β_k are the shape parameters of the k-th Beta distribution. In all models, we set the number of beta components to $K = 10$.

Linear BMM. For the Linear BMM model, the input consisted of 12 features made up of the position of each obstacle, the orientation of each obstacle, and one hot

encodings of the drop hole. The linear model architecture consisted of a Layer Normalization on the 12 input features, a dense linear matrix with 2048 output units and a subsequent Dropout layer on the 2048 units before the the BMM output projection heads.

MLP BMM. Similarly to the Linear BMM, the input to the MLP consisted of 12 input features made from the position and orientation of each obstacle and a one-hot encoding of the drop hole. The first layer of the multilayer perceptron (MLP) was defined similarly to the Linear BMM using a dense layer of 2048 units with the Layer Norm before and Dropout after. We apply a GELU nonlinearity (Hendrycks & Gimpel, 2016) after the dropout and add another dense layer block of 2048 units before the BMM output projection heads.

CNN BMM. Lastly, to model bounded human response distributions with high flexibility and smoothness, we implemented a BMM whose parameters were predicted by a deep residual convolutional neural network (CNN) (He et al., 2016). Instead of using a feature based input, the CNN architecture receives an image which is encoded as a three-dimensional input tensor (e.g., channels $\times$ height $\times$ width) and processes it through a series of residual convolutional blocks, followed by adaptive pooling and dense projection heads for parameter estimation.

The architecture begins with an initial 7×7 convolution with stride 2 and padding 3, followed by batch normalization, GELU activation, and optional Gaussian noise injection. This preprocessing layer increases representational capacity while reducing spatial resolution.

Following the input stage, the model includes three residual stages defined by each with 2 residual blocks and each with channel sizes 12, 24, and 48 respectively. The first block in each stage performs downsampling via stride 2 (except the first stage, which retains stride 1), and increases the number of channels. Residual connections use 1×1 convolutions when needed to match channel dimensions.

Each residual block applies the following operations:

1. batch normalization
2. 0 centered Gaussian noise with $\sigma = 0.01$,
3. Two 3×3 convolutions (with padding 1 and no bias),
4. GELU activation between and after convolutions,
5. Residual skip connection and final GELU activation.

Following the final residual block, the output feature map is pooled to shape $B \times C \times 1 \times 1$ via adaptive average pooling, then reshaped and passed to the BMM output projection heads.

Training Procedure. Each model was pretrained on data consisting of deterministic physics simulation outcomes and subsequently trained to maximize the log-likelihood of observed human responses y under the predicted mixture distribution:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} c_k(x) \cdot \text{Beta}(y \mid \alpha_k(x), \beta_k(x)) \right)$$

Numerical stability was ensured by computing log-sum-exp expressions and using appropriate regularization. Optimization was performed using the Adam optimizer, with learning rate, hidden dimensions, dropout rate, and number of mixture components selected via cross-validation.

Dataset and Splits. The pretraining physics simulation data consisted of 50,000 generated image output pairs created from the Chipmunk physics engine. In each data sample, the models trained on results from all three drop holes. After pre-training we performed a cross validation over 50/50 splits of the human data. We partitioned the human data into splits of half training and half validation data consisting of trial-level response data from human participants across 40 distinct environments ('worlds'). Each world contained a set of target locations ('holes') and corresponding human response distributions. We report the average validation performance over these splits.

Appendix B

Sample Prediction Task Trials with Model Predictions

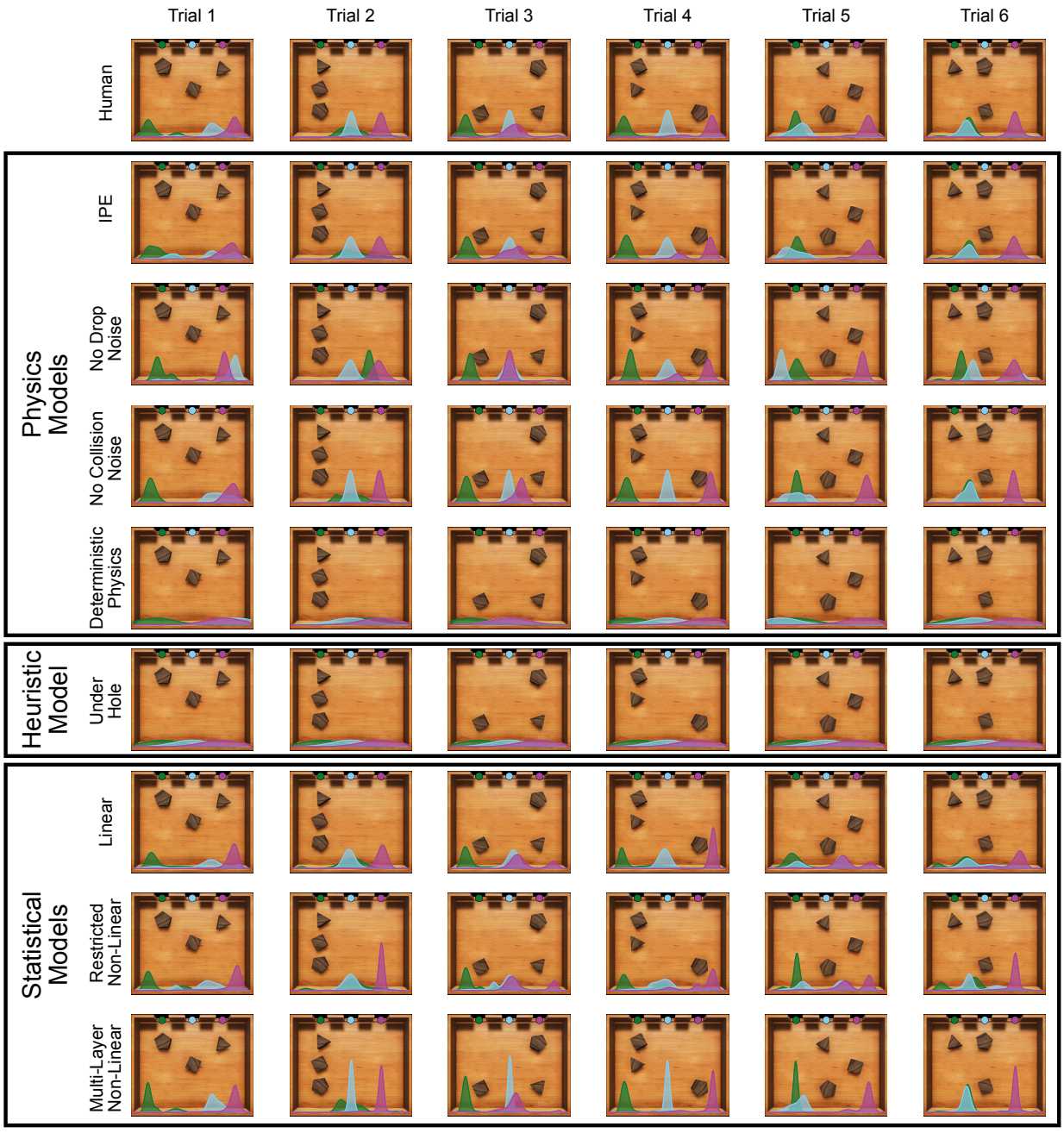

Figure B1

Sample Prediction Trials with Human distirbutions and model predictions. Models are organized as physical models (IPE and lesions), the heuristic under hole model, and the statistical models.