A nonparametric model of object discovery

Pat C Little (pat.little@nyu.edu)
Todd M Gureckis (todd.gureckis@nyu.edu)

Department of Psychology, New York University
6 Washington Place, New York, NY 10003

Abstract

A detailed model of the outside world is an essential ingredient of human cognition, enabling us to navigate, form goals, execute plans, and avoid danger. Critically, these world models are flexible—they can arbitrarily expand to introduce previously-undetected objects when new information suggests their presence. Although the number of possible undetected objects is theoretically infinite, people rapidly and accurately infer unseen objects in everyday situations. How? Here we investigate one approach to characterizing this behavior—as nonparametric clustering over low-level cues—and report preliminary results comparing a computational model to human physical inferences from real-world video.

Keywords: intuitive physics; causal perception; object discovery; physical reasoning

Introduction

Extracting meaning from the noisy data provided by our senses requires an inferential leap from sensations to understanding, classically modeled as a process of Bayesian inference (Helmholtz & Southall, 1925; Kersten, Mamassian, & Yuille, 2004; Mansinghka, Kulkarni, Perov, & Tenenbaum, 2013). The goal of this inference is a coherent and generally accurate model of the world, one that can support planning (Craik, 1943), physical reasoning (Gentner & Stevens, 2014), and logical inference (Johnson-Laird, 1980). The challenge is that building an accurate model of our environment is not always straightforward. Even in the best case, where all the relevant objects are in plain view, we must still identify them across the infinite variations in how they could appear to our senses (Biederman, 1987). Often, the picture is even bleaker: whether by shadow, occlusion, or haze, some relevant objects are not visible at all, and can only be inferred.

Clever experiments like those depicted in Figure 1 have shown how we can effortlessly determine a great deal about even hidden or invisible objects. In these examples, the objects or agents are not sensed directly, and yet we recognize them by the physical effects they have on the world around them. Such inferences are not limited to the laboratory, but happen regularly in everyday life: when your silverware drawer won't close, you can infer that a misplaced spoon is in the way.

These examples highlight two critical properties of our world models: first, rather than being limited to a fixed set of known objects, they are *expandable*, able to accommodate additional hidden entities when required in order to account for new observations. Second, the information used to support such inferences can make use of physical intuitions (e.g., of the way cloth drapes, or the way objects fall). How do these properties enable us to discover objects?

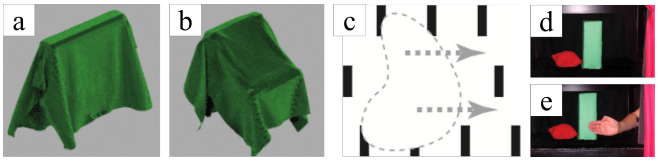

Figure 1: Selection of stimuli from previous studies of object inference. (a and b) Only a cloth is visible in each panel, yet it is clear that two other objects are present, and that one is a chair (Yildirim et al., 2016). (c) An object the same color as its background can nevertheless be clearly perceived when it moves in front of other objects, occluding and revealing them in turn (Palmer et al., 2006). (d and e) After seeing a beanbag fly over a wall from the right, even infants infer that an agent was responsible, and expect to see a hand emerge on the right side rather than the left (Saxe et al., 2005).

Hidden object discovery as latent causal inference

To discover a hidden object is to infer an explanation—a latent cause of the data we observe (Shams & Beierholm, 2010). "What object could cause the particular draping in Figure 1B?" we might ask, and propose a chair as the answer. Our physical reasoning abilities let us generate candidate causes and predict their likely effects, yielding plausible explanations for otherwise puzzling observations (Yildirim et al., 2016; Gerstenberg, Siegel, & Tenenbaum, 2021).

In particular, we generate new explanations when our current models can't explain the data we observe. We know cloth doesn't support itself, and that inanimate objects don't randomly disappear or propel themselves over walls, so when they seem to do so, we decide that something else must be behind this behavior and start to seek out a suitable explanation (Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018).

While a great deal of progress has been made in recent years toward understanding human inference in physical reasoning (Kubricht, Holyoak, & Lu, 2017), such efforts have generally focused on prediction and inference over already-known objects. Even models that infer latent *properties* like mass and friction (Battaglia, Hamrick, & Tenenbaum, 2013;

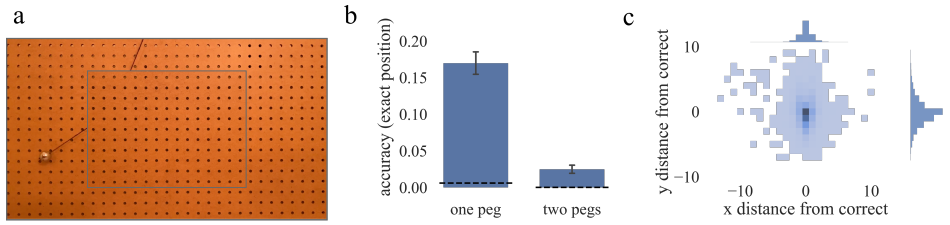

Figure 2: (a) The experimental setup. On each trial, a video of an actual physical pendulum looped continuously as participants selected the holes that they thought contained pegs. (b) Proportion of trials where the precisely correct peg(s) were selected, split into trials with one peg or two. Dashed lines show chance level, error bars are bootstrapped 95% confidence intervals. (c) Heat map showing distance between each response and the correct answer, which is centered at (0,0).

Ullman et al., 2018) do not generally address the discovery of new *objects* (though see Carroll & Kemp, 2015).

On what basis can people infer such latent causes? One way would be to mentally simulate all the possible causes and select the one whose simulated consequences best match the data actually observed (Yildirim, Siegel, Soltani, Ray Chaudhuri, & Tenenbaum, 2024). Situations with a very large number of possible causes, though, might be intractable without more bottom-up processing. Here we propose one possible strategy: When one solid object interacts with another, the motion of each object is generally different before, during, and after the interaction. We suggest that these natural boundaries may form distinct "clusters" in time that the mind can leverage to make inferences about the parameters of the interaction.

We assess the viability of this idea in two ways: first we report a behavioral experiment where participants had to infer the number and position of hidden objects on the basis of their interactions with visible objects. Second we propose a novel computational model that takes as input raw video of the motion of real-world objects and infer the number and properties of latent causes in a scene. We compare multiple variants of the model that rely on distinct low-level cues in order to better understand this cognitive ability.

Experiment

For our experiment we recorded videos of actual physical scenes. We then occluded portions of the video and asked participants to "fill in" the missing aspects of the scene (related to a popular task in computer vision known as "inpainting"). The only basis for filling in these details is to infer something about how the occluded objects might influence the behavior of visible ones. Our goal was to elicit inferences of hidden objects that were easy enough to tap into our everyday physical reasoning abilities, but not so easy that participants would be at ceiling. In addition, we wanted the method of responding to admit to easy quantitative analysis.

Methods

Participants We recruited 450 participants from Prolific, who were paid $2 each for participating in the experiment (which took approximately 8 minutes). Participants were excluded if they did not complete the task and correctly answer both catch trials (see below), or if they responded implausibly quickly (within four seconds), leaving N=367 after exclusions. We did not exclude for poor performance on the main task, as establishing baseline human performance was one aim of the experiment.

Stimuli Participants watched 1080p, 30-frame-per-second videos of a real physical pendulum swinging in front of a wooden peg board (see Figure 2a). A peg in the top middle of the board, out of frame of the camera, served as an anchor—one end of a piece of black string was attached to the anchor peg, and the other end to a steel ball approximately 2 cm in diameter. The length from the anchor peg to the center of the ball was approximately 50 cm. A central rectangular region 15 holes wide and 11 holes tall was photographed and later digitally superimposed on each video to act as an occluder. Pegs were placed within the occluding region but never in the holes along its edge (where their shadows would have been visible). This left 117 distinct peg placement locations. Each video had either one or two pegs in the occluded region. There were 117 one-peg videos (one for each peg location) and 117 two-peg videos (sampled randomly from the $\binom{117}{2} = 6786$ arrangements of two pegs) for a total of 234 videos. In each video, the pendulum was set swinging by hand such that the string contacted the occluded pegs during part of its motion. Finally, the videos were trimmed to show one complete period of the pendulum (approximately one second), so that they could seamlessly loop for the duration of each trial.

Task Participants watched the videos and selected the hole(s) which they thought contained pegs. A demo of the task as shown to participants can be accessed here. The task took approximately 8 minutes and participants were paid $2, with a potential $1 bonus for performance. After an instruction phase, participants were shown videos like the one depicted in Figure 2a and selected peg locations by clicking on the corresponding hole (without time pressure). Clicking would superimpose an image of a placeholder peg, and clicking again would remove the placeholder. In this way, participants could refine their estimates of the peg location(s) before submitting the trial. Each participant completed 26 trials, in addition to two catch trials where the occluder was absent (so the task became trivial—"click on the visible peg").

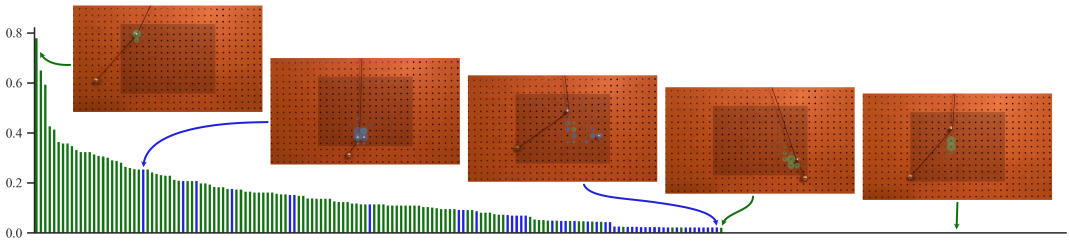

Figure 3: Mean accuracy across videos, sorted. Green bars represent videos with only one peg, blue bars represent videos with two pegs. Heat maps and still frames are shown for some example videos. The slightly darker region within each still frame shows the region that was occluded to participants. The string and ball are visible in these examples but were occluded within the dark region for subjects. Superimposed on the videos are colored circles, where size of each circle is proportional to the number of participants who selected that hole. Green circles represent participants who selected only one hole on that video and blue circles represent participants who selected two holes.

Results

A summary of results for all 367 included participants is shown in Figure 2. While absolute accuracy was low (see Figure 2b), success on a trial required selecting the one correct location out of 165 options—a high bar, and made even higher when there were two pegs. Further, since participants did not know the number of pegs present, the the number of possible answers is—in principle—$2^{165} \approx 5 \times 10^{49}$. Participants therefore performed reliably above chance, and as can be seen in Figure 2c, even their incorrect responses were tightly clustered around the correct position, with slightly more variance in the y dimension than the x.

Figure 3 shows the striking variation in difficulty across videos and the patterns of errors that participants made. Videos with only one peg naturally had much higher accuracy overall (selecting one peg instead of two means half as many opportunities for error), and the two-peg videos with greater accuracy tended to feature pegs near the bottom or sides of the occluding region.

Some patterns of performance are less obvious. For example, the highest-accuracy video (far left of Figure 3) features a single peg in the upper left corner of the occluded area, not that far from the peg position in one of the videos which no one answered correctly (far right of Figure 3). Across all the videos, many features contribute to difficulty including distance from the edge of the occluded region and time the string spends in contact with the peg. One goal for our computational model is to capture some of what accounts for the variation in human difficulty across videos.

Computational Model

Participants in our experiment detected unseen objects reliably above chance, despite the large number of possibilities that the task admitted. At the same time there was substantial variation in performance for different stimuli. How might the mind make reasonable inferences in so large a hypothesis space and what might explain these variations in performance?

We propose that while watching a scene, people estimate cues, derived in part from their physical understanding of the world, and track the evolution of those cues over time. Clustering of these cues in time provides evidence for the existence of latent causes (in this case, objects). To accomplish this clustering, our approach leverages the power of Bayesian nonparametric models (Anderson, 1991; Sanborn, Griffiths, & Navarro, 2006; Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006; Gershman, Blei, & Niv, 2010; Gershman, Norman, & Niv, 2015). A critical feature of these models is that they have no fixed capacity. Where a fixed-capacity model might consider n possible causes for each new observation, a nonparametric model always reserves some probability that none of the previously considered causes is the right one, and that a new cause must be added. In this way, such models can grow in complexity just as much as the data demand.

In the following section we describe the input to the model, the model structure, the inference procedure, and then turn to results (see Figure 4 for an overview).

Preprocessing

The first step in applying the model requires extracting lower-dimensional representations of the video frames. In particular, we are interested in structural features like the ball position and string angle, which are relevant for estimating the ultimate quantity of interest (peg positions). Building our inferences up from raw video is inherently noisy, but we see this noise as a virtue of our approach.

As it is grounded in real-world video, the model works with data that has similar noise characteristics to what humans face. By contrast, much of the work on human intuitive physics employs artificial renderings of objects in simulation software (Battaglia et al., 2013; Smith, Battaglia, & Vul, 2013; Ullman et al., 2018). In physics simulators, the "ground truth" position of each object can usually be read out directly and passed into an inference model (perhaps with added perceptual noise). Here, our estimates of the properties of the scene are limited to what can be extracted from a video, and the model has to make do with the same im-

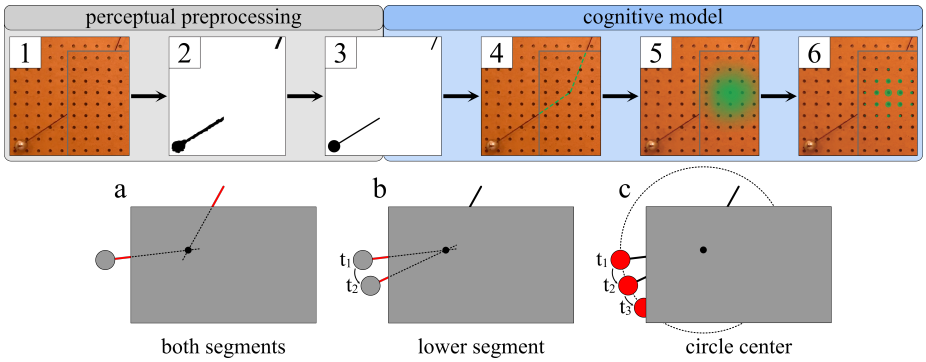

Figure 4: (Top) Schematic of processing steps in the model. (1) Raw video frame. (2) Ball and string isolated from the background. (3) Idealized ball and string fit to isolated region. (4) Cue applied to evaluate likely peg positions (here, the "both segments" cue). (5) Peg location distribution refined over frames. (6) Discrete response location sampled from the location distribution. (Bottom) The three cues to peg position evaluated under the model. (a) Extrapolating the both visible segments of the string. (b) Extrapolating only the lower segment of string. (c) Identifying the center of the circle traced out by the ball.

perfect recordings (featuring glare, lens distortion, etc.) that the participants saw. Similarly, the physical dynamics of our videos necessarily include all the nuisance factors (e.g., friction, slipping, air currents, manufacturing imperfections) that are often ignored by physics simulators.

Here, we preprocessed the original video files to extract the positions of the ball and visible string segments in each frame (see Figure 4 for examples at each stage). We first used the Segment Anything model (Kirillov et al., 2023) along with morphological image processing operations to remove remnants of the background image. This left reasonably accurate masks of the ball and string, which were then fit with a simple model of a disk and two string segments using the SciPy implementation of Powell's method (Powell, 1964; Virtanen et al., 2020). Finally, we superimposed the best fitting parameterizations on the videos and, where any discrepancies were noticeable, adjusted the parameters to better match the video (approximately 10% of the videos had at least one frame adjusted by hand). An example of a fitted frame is shown in Fig 4-3.

Low-level cues

After the preprocessing steps (1-3 in Figure 4), the model is equipped with estimates of the ball and string positions. At this point an ideal rational analysis for this problem could make use of a huge number of regularities in a video (cues) to infer the peg position. For example, because the period of a pendulum depends on its length, one could use the time it takes for the pendulum to complete a portion of its swing to constrain the possible peg positions, even if the ball is hidden for a portion of that time. Such a cue would be difficult for a human to compute, however, and so it is unlikely to play a major role in human performance. Other cues, though, are easier to compute—perhaps even in real time by low-level perceptual processes. Here we focus on this latter sort of cue, which we feel are a more natural substrate for a model of everyday physical inferences.

Foundational work in visual perception has investigated the complex ways we extrapolate contours into sensible shapes (Kellman & Shipley, 1991) and revealed how the visual system makes use of detailed kinematic cues, which can even be sophisticated enough to discover invisible objects, as in Figure 1C (Palmer et al., 2006). We suggest that a similar cue might be at work here. For example, people might assume that a string under tension will be straight unless it comes into contact with a solid object, such that any misalignments in the visible segments of string are evidence of an object at their intersection point (Figure 4-a). A cue like this could be implemented as a conscious strategy but could also reflect more basic principles of perceptual processing, like good continuation.

The fact that a bending string implies contact with a surface is only one way to use physics to infer the presence of an object—here we also consider two more cues. We can find the point of intersection using only the lower string segment and integrating across multiple frames (Figure 4-b). Lastly, we can infer that a peg is located at the center of the imaginary circle traced out by the ball as it swings (Figure 4-c; note that for a peg with radius greater than zero, the actual path of the ball is not a perfect circle but an involute of a circle).

Each of these is a physically accurate and computationally plausible method to infer the hidden pegs, but some may be more reliable under real world noise conditions, such that reliance on different cues might lead to different patterns of behavior. One goal of our model is to approximate the entire inferential process—taking in real world video data and operating over that input according to distinct cues, to find if any are more likely than others.

Nonparametric model

Following (Gershman et al., 2010), we treat the problem of attributing different observed cues to different latent causes as a clustering problem over an unknown number of clusters, and apply a Dirichlet process mixture model. Interestingly,

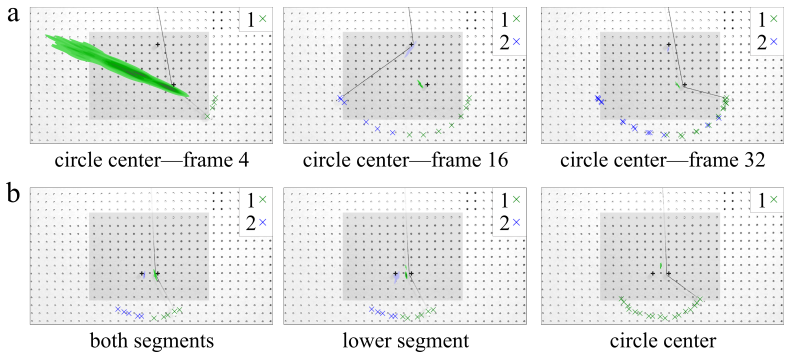

Figure 5: Model inferences: blobs indicate likely peg positions under each cluster, and x's indicate the cluster assignment of the observation that happened when the ball was at that position. (a) Cluster likelihood and assigned observation partitions for different frame numbers as a video unfolds. (b) Final likelihoods (faint) and partitions from the same video for the three cues.

the design of our task very naturally aligns with the structure of a mixture model. At each point in time, the string is in contact with one or more pegs, and the peg farthest down the string acts as a pivot point of a pendulum. As a result, the entire system evolves as a mixture of pendula with different lengths. For example, a pendulum might swing around the top peg for some part of its overall period and then, when the string hits another peg, it begins to act as a new, shorter pendulum for a while, before returning to swing around the top peg. The goal of the model then is to partition the video frames based on their different latent causes—here, the hidden pegs.

Table 1: Model parameters

Parameter	Description	Free?
α	cluster concentration	Free
λ	lapse rate	Free
σ_R	response error	Free
σ_O	observation error	Fixed
(μ_x, μ_y, μ_r)	peg prior means	Fixed
$(\sigma_x, \sigma_y, \sigma_r)$	peg prior sds	Fixed

Model inference

We estimate the model using a particle filter. The particle filter consists of m particles, which each maintain a running partition $\boldsymbol{c}$ of the t video frames presented so far, assigning them to as many as t different clusters. For example, after six frames, one particle might have $\boldsymbol{c} = [1,1,1,2,2,1]$—that is, it has assigned frames 1, 2, 3, and 6 to one cluster, and frames 4 and 5 to a second cluster. See Figure 5 for examples of frames assigned in this way.

The first frame is always assigned to the first cluster, i.e., when $t = 1$, all m particles are set to $\boldsymbol{c} = [1]$. On each subsequent frame[1], the prior probability that frame i will be assigned to cluster k is given by

$$P(c_i = k) = \begin{cases} \frac{N_k}{i+\alpha} & \text{if k is old} \\ \frac{\alpha}{i+\alpha} & \text{if k is new,} \end{cases}$$

where N_k is the number of observations already assigned to cluster k, and concentration parameter α reflects a prior belief about the distribution of clusters. When $\alpha = 0$, all frames are assigned to cluster 1, and when $\alpha = \infty$, each frame is assigned to a different cluster.

The particles are weighted and resampled (with replacement) after each frame, so that the weight $w_{l,t}$ of particle l at time t is proportional to the likelihood of observation o_t given the particle's cluster assignments and the previous observations:

$$w_{l,t} \propto P(o_t | \boldsymbol{c}_{l,t}, \boldsymbol{o}_{1:t-1}),$$

where $\boldsymbol{c}_{l,t}$ is the vector of cluster assignments of particle l at time t.

Before any observations have been assigned to it, each cluster begins with a Gaussian likelihood over the three dimensions x, y, and r, where (x,y) is the position of the peg. For the "both segments" and "lower segment" cues, r is the radius of the peg itself, and for the "circle center" cue, r is the radius of the near-circle traced out by the ball as it makes contact with that peg. Informally, the likelihood of a given observation O is how "compatible" it is with each possible value of (x,y,r), weighted by the value of the likelihood at that point. For the circle center cue, that is:

$$P(o_t | \boldsymbol{c}_{l,t}, \boldsymbol{o}_{1:t-1}) \propto \int_{x,y,r} P(x,y,r) \frac{1}{\sigma_O} \varphi\left(\frac{||o_t - (x,y)||_2 - r}{\sigma_O}\right),$$

where $\varphi(\cdot)$ is the probability density function of the standard Gaussian distribution, $||\cdot||_2$ is the Euclidean norm, and σ_O is the observation error, a free parameter of the model which acts as a sort of Gaussian kernel. See Figure 5 for plotted examples of likelihoods.

The likelihoods for the segment cues are defined similarly, where the distance measure is the distance from the

[1]Note that the model only updates on frames where the ball or string segment are visible—on some frames, they pass behind the occluder entirely. Because the ball is the last part to disappear behind the occluder, the circle center model is given more observations over which to find the peg (see the colored x's in Figure 5).

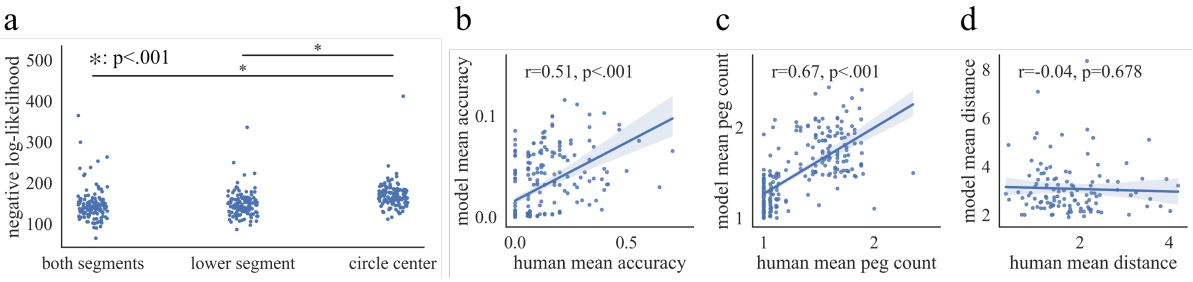

Figure 6: (a) Comparison of participant-level fits under each of the three cues. Each dot represents a single participant. Asterisks indicate a difference under Tukey's honest significance test. (b, c, d) Comparison of lower segment model and human results. Each dot represents the mean for a single video. Distances (d) only for videos with one peg. Human means are computed across participants, and model means across 20,000 samples from 10 runs (see text).

line(s) traced by the segments. Observations are assigned to clusters in proportion to their posterior probabilities, and once assigned, the values of the peg location likelihoods for each cluster are re-weighted based on how compatible they are with the new observation. In practice, we approximate these likelihoods with another particle filter, sampling (x, y, r) points from the prior, re-weighting them, and resampling them (in proportion to $P(x, y, r|\boldsymbol{O}_{1:t})$ with each new observation.

Modeling guesses about peg location

To sample a response (a guess as to the location of each peg in the video) from the model, we generate a peg location from each cluster one at a time. First, we find the likelihood of each grid position under the first cluster, applying another Gaussian kernel governed by a response error parameter, σ_R, to form a probability mass function. Next we sample a peg position in proportion to these probabilities, and repeat for each remaining cluster. Another model parameter, lapse rate λ, determines the (independent) probability that each cluster will be skipped during this process. To evaluate the likelihood of a response, we calculate its likelihood under the probability mass functions of the clusters, accounting for the lapse rate. Because the integrals are approximated and the cluster assignments are probabilistic, the model will provide slightly different samples for different random seeds. To average out this variation, we repeat the response sampling procedure for 20 model runs, and take the average likelihood of responses over the runs.

Preliminary model evaluation

To evaluate how the three different cues (both segments, lower segment, and circle center) compare under the model, we fit the model independently using each of them. For all these evaluations, we chose a random sample of 130 participants (approximately 1/3 after exclusions), to reduce the computational cost. For this preliminary evaluation the model had three free individual-level parameters, α (clustering probability), λ (lapse rate), and σ_R (response error), holding the other parameters fixed. Using Optuna (Akiba, Sano, Yanase, Ohta, & Koyama, 2019), we efficiently sampled parameter values, ran the model for each of the videos that a single participant had seen, and evaluated the likelihood of that participant's responses under the model. The parameter values with the highest likelihoods (lowest negative log likelihoods) were kept, and the distributions of negative log likelihoods were compared (see Figure 6a). In brief, the models using the "both segments" and "lower segment" cues outperformed the model using the "circle center" cue (the different cue models had the same number of parameters, so negative log likelihoods are an appropriate metric for comparison).

As an additional check, we also simulated responses to see how well the model was able to capture some other aspects of human behavior. We chose one of the two better-performing models (the lower segment model) and used the participant-level maximum-likelihood parameter values to sample 20,000 responses, 20 times per participant, for each of the 130 participants separately. Then we measured the model's accuracy (did they select the exact correct peg?), peg count, and distance from the correct peg. A comparison to the relevant human data shown in Figure 6b-d. While the model accuracy was lower overall, it tended to perform better on those videos where participants also had higher accuracy. Comparing the distance between the guessed and correct peg positions on the one-peg videos, however, highlights a limitation: while the model can explain some of the variance between one-peg and two-peg videos, it does not well explain the variance within the one-peg videos (Figure 6d).

Discussion

Here we described a flexible method for modeling object discovery, combining nonparametric models with simple perceptual cues to infer latent causes in an intuitive physical reasoning task. In this work, both our model and human participants made judgments on the basis of raw, real-world video, with all the associated noise and imperfections that are not typically present in idealized physical simulations. Future work will evaluate the model on a wider variety of physical inference tasks (such as inferring hidden surfaces from collisions) and directly compare the present model with alternatives grounded in mental simulation.

Acknowledgments

The authors would like to thank Emily Liquin and members of the NYU Computation and Cognition Laboratory for their valuable feedback. Support provided by NSF BCS-2121102 awarded to T.M.G.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2623–2631).

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, *98*(3), 409.

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*, 18327–18332.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, *94*(2), 115.

Carroll, C. D., & Kemp, C. (2015). Evaluating the inverse reasoning account of object discovery. *Cognition*, *139*, 130–153.

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

Gentner, D., & Stevens, A. L. (2014). *Mental models*. Psychology Press.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological review*, *117*, 197.

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

Gerstenberg, T., Siegel, M., & Tenenbaum, J. (2021). What happened? reconstructing the past through vision and sound.

Helmholtz, H. v., & Southall, J. (1925). *Helmholtz's treatise on physiological optics* (Vol. 3). Optical Society of America.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive science*, *4*(1), 71–115.

Kellman, P. J., & Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive psychology*, *23*(2), 141–221.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Aaai* (Vol. 3, p. 5).

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annu. Rev. Psychol.*, *55*, 271–304.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... others (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

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

Mansinghka, V. K., Kulkarni, T. D., Perov, Y. N., & Tenenbaum, J. (2013). Approximate bayesian image interpretation using generative probabilistic graphics programs. *Advances in Neural Information Processing Systems*, *26*.

Palmer, E. M., Kellman, P. J., & Shipley, T. F. (2006). A theory of dynamic occluded and illusory object perception. *Journal of Experimental Psychology: General*, *135*, 513.

Powell, M. J. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, *7*(2), 155–162.

Sanborn, A., Griffiths, T., & Navarro, D. (2006). A more rational model of categorization.

Saxe, R., Tenenbaum, J., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10-and 12-month-old infants. *Psychological Science*, *16*(12), 995–1001.

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

Smith, K. A., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. *Proceedings of the annual meeting of the cognitive science society*, *35*.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive psychology*, *104*, 57–82.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... others (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, *17*(3), 261–272.

Yildirim, I., Siegel, M. H., Soltani, A. A., Ray Chaudhuri, S., & Tenenbaum, J. B. (2024). Perception of 3d shape integrates intuitive physics and analysis-by-synthesis. *Nature Human Behaviour*, *8*(2), 320–335.

Yildirim, I., Siegel, M. H., & Tenenbaum, J. B. (2016). Perceiving fully occluded objects via physical simulation. *Proceedings of the annual meeting of the cognitive science society*, *38*.