Inferring the existence of objects from their physical interactions

Pat C Little (pat.little@nyu.edu) and Todd M Gureckis (todd.gureckis@nyu.edu)
Department of Psychology, New York University
6 Washington Place, New York, NY 10003

Abstract

A fully occluded object cannot be perceived directly, but we can still infer its existence from the effect it has on the motion and behavior of other, visible objects. Here we report the results of a behavioral experiment designed to elicit these sorts of inferences and quantify their reliability. Our experiment leverages videos of real-world objects interacting under real-world physics (specifically, interrupted pendulum motion). We propose a preliminary model for how the mind might efficiently infer the position and number of occluded objects simply from the effect they have on the visible physics of a scene.

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

Introduction

In order to navigate an environment full of occlusions, shadows, and overlapping sounds, the mind must integrate noisy observations with a great deal of prior knowledge—possibly using a form of Bayesian inference (Kersten, Mamassian, & Yuille, 2004; Mansinghka, Kulkarni, Perov, & Tenenbaum, 2013; Pouget, Beck, Ma, & Latham, 2013). An important tool enabling this inference is our ability to construct a detailed physical model of the objects around us and predict their properties and future behavior (Battaglia, Hamrick, & Tenenbaum, 2013; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018).

Perhaps even more impressive than accurately extrapolating from noisy observations of visible objects is our ability to infer the existence of objects that we *cannot directly perceive at all*. Consider a pedestrian who, while putting on her slippers, notices her neighbor slipping and sliding on the sidewalk. *There must be ice on the ground*, she realizes, and so reaches for her boots instead. More than simply predicting the behavior of objects in view, she is able to discover a hidden substance solely on the basis of its physical interactions (Carroll & Kemp, 2015).

We sought to bring these sorts of inferences into the laboratory with a novel behavioral task in which participants inferred the existence and position of an unknown number of hidden objects from a video of a real-world scene. Making such inferences required them to integrate perceptual information over time and then generate plausible hypotheses (Dasgupta, Schulz, & Gershman, 2017) to ultimately determine the true latent properties of the scene. In addition to our behavioral results, we describe a preliminary model for how an appropriate hypothesis might be efficiently selected.

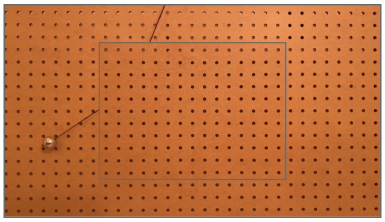

Figure 1: The experimental setup. Each video looped continuously as participants made their response.

Methods

Behavioral Experiment

Stimuli were 1080p videos of a pendulum swinging in front of a partially occluded pegboard (see Figure 1). In each video, either one or two pegs were inserted into the board behind an occluding rectangle such that they were not directly visible but still made contact with the string during part of its motion. Participants (N=450) recruited from Prolific were asked to select the holes which they thought contained pegs.

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 1, and selected peg locations by clicking on the corresponding hole in the pegboard (without time pressure). Each participant completed 26 trials, in addition to two catch trials where the occluder was absent (so the task became trivial). Participants were excluded if they did not complete the task and correctly answer both catch trials, or if they responded too quickly (within four seconds), leaving N=367 after exclusions.

Computational Model

Visual processing makes use of sophisticated kinematic cues (Palmer, Kellman, & Shipley, 2006), which may even involve limited forms of physical inference (Firestone & Scholl, 2017; Little & Firestone, 2021). Our model assumes that a similar perceptual cue is at work when participants observe scenes like that shown in Figure 1. In particular, we suggest that the two visible segments of string are extrapolated to find their point of intersection. This “where would the lines meet?” cue is imperfect (the string bends slightly around the pegs, for example), but it is sufficient to approximate the true peg location. Future iterations of this model will compare alternative cues (e.g., the center of the arc traced out by the path of the ball).

Extrapolating the string segments gives a reasonable guess for each of the frames in which the ball was visible and the line segments non-colinear, but these guesses do not by themselves determine the number and location of distinct pegs.

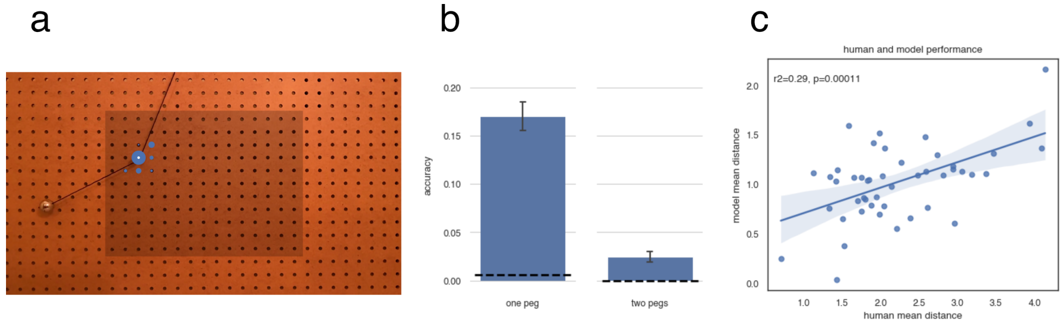

Figure 2: (a) Heat map of participants' responses for the trial depicted in Figure 1. (b) Comparison of model and human performance. Each dot indicates performance for a single video. Distances are scaled so that two adjacent holes in the pegboard are one unit apart. Human means are computed across participants, and model means across 100 runs. (c) 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 95% CI.

The relevant physical dynamics of the ball and string are subtle (the length of the lower string segment, for example, changes slightly as the string wraps around a peg), but we can simplify the situation substantially by treating it as a clustering problem over an unknown number of clusters: given a string-extrapolation point from each frame, what are the clusters of points, and where? A natural choice for this type of problem is a Dirichlet process mixture model, which has a history of successful applications in category learning (Griffiths, Navarro, & Sanborn, 2006; Gershman, Blei, & Niv, 2010). Importantly, these models are *non-parametric*, in that the number of clusters does not need to be specified in advance.

In our case, we first extrapolated the line segments on each frame, adding perceptual noise in proportion to the length of the visible line (note: this is an imperfect approximation; see Morgan, 1999). This yielded an estimated peg position for each eligible frame. We then approximated the posterior of the infinite Gaussian mixture, with each cluster mean representing an inferred peg position. To compare these continuous inferences to human responses (which were discretized by the pegboard grid), we first assigned each of the model's inferred cluster means to the nearest valid peg location, and then calculated the mean distance between these and the true peg locations for both the model and the participants.

Results

Participants performed reliably above chance (see Figure 2b). While absolute accuracy was low, 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. The results for the 40 participants (after exclusions) who completed one example trial are shown in Figure 2a, where the larger circles are locations that participants selected more often. We also measured participants' mean distance from the correct peg location(s), scaled so that selecting the peg immediately adjacent to the correct peg would be a distance of one. The correlation between human and model performance on this measure is depicted in Figure 2c. This preliminary model fit suggests that we are capturing some aspects of the variation in difficulty between trials, but a more thorough model evaluation is needed before any substantive conclusions can be drawn.

Discussion

Object discovery can take many forms, from recognizing an object covered by a cloth (Yildirim, Siegel, & Tenenbaum, 2016) to discovering a previously-unknown planet (Galle, 1846). Our work describes an simple and intuitive paradigm for examining these types of inductive causal inferences. An important feature of this project is the use of videos of real-world physical events. In contrast, much of the work in the psychology of 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 (with added perceptual noise). Here, our estimates of the properties of the scene were limited to what could be extracted from a video. This keeps us honest—our model has to make do with the same imperfect recordings that the participants saw. Similarly, the physical dynamics of our videos necessarily included all the nuisance factors (e.g., friction, slipping, air currents, manufacturing imperfections) that are often ignored by physics simulators.

Inferring the existence of hidden objects is difficult because of the sheer number of latent causes that could plausibly be operating in our physical environments. Considering all these possibilities—the problem of hypothesis generation—can be the primary computational challenge for making accurate inferences (Dasgupta et al., 2017; Phillips, Morris, & Cushman, 2019). Our approach leveraging Bayesian non-parametric models enables us to explain how people flexibly adapt the complexity of their hypothesis to the structure of the physical scene.

Acknowledgments

Support provided by NSF award 2121102 to TMG.

References

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.

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

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive psychology*, *96*, 1–25.

Firestone, C., & Scholl, B. (2017). Seeing physics in the blink of an eye. *Journal of Vision*, *17*, 203–203.

Galle, J. (1846). Account of the discovery of the planet of Le Verrier at Berlin. *Monthly Notices of the Royal Astronomical Society*, *7*, 153.

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

Griffiths, T. L., Navarro, D. J., & Sanborn, A. N. (2006). A more rational model of categorization. *Proceedings of the annual meeting of the cognitive science society*, *28*.

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

Little, P. C., & Firestone, C. (2021). Physically implied surfaces. *Psychological Science*, *32*, 799–808.

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*.

Morgan, M. (1999). The poggendorff illusion: A bias in the estimation of the orientation of virtual lines by second-stage filters. *Vision research*, *39*, 2361–2380.

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.

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

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, *16*, 1170–1178.

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.

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*.