Efficient Partial Simulation Quantitatively Explains Deviations from Optimal Physical Predictions

Ilona Bass
Department of Psychology
Harvard University
ibass@fas.harvard.edu

Kevin Smith
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
k2smith@mit.edu

Elizabeth Bonawitz
Graduate School of Education
Harvard University
elizabeth_bonawitz@gse.harvard.edu

Tomer Ullman
Department of Psychology
Harvard University
tullman@fas.harvard.edu

Abstract

Humans are adept at planning actions in real-time dynamic physical environments. Machine intelligence struggles with this task, and one cause is that running simulators of complex, real-world environments is computationally expensive. Yet recent accounts suggest that humans use mental simulation in order to make intuitive physical judgments. How is human physical reasoning so accurate, while maintaining computational tractability? We suggest that human behavior is well described by *partial simulation*, which moves forward in time only parts of the world deemed relevant. We take as a case study Ludwin-Peery, Bramley, Davis, and Gureckis (2020), in which a *conjunction fallacy* was found in the domain of intuitive physics. This phenomenon is difficult to explain with full simulation, but we show it can be quantitatively accounted for with partial simulation. We discuss how AI research could make use of efficient partial simulation in implementations of commonsense physical reasoning.

1 Introduction

Humans are quite good at reasoning about physical events in their environments. We implicitly make predictions about objects' properties (Leslie, Xu, Tremoulet, & Scholl, 1998), and how they will interact (Kominsky et al., 2017), starting as early as infancy (Baillargeon, 2004; Spelke, Breinlinger, Macomber, & Jacobson, 1992). Yet our current state-of-the-art engineered systems come nowhere near human-level competence on these tasks – they are often less accurate than people, or require extensive computation that makes real-time prediction and control intractable. Thus we propose looking to human physical prediction to understand useful shortcuts to design into our AI systems.

Some have suggested that humans possess an *intuitive physics engine*, similar in structure to what is used to create physics in modern video games (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016). Using such a model, we can run forward simulations of possible outcomes, allowing us to make predictions about the probabilities of physical events (Ullman, Spelke, Battaglia, & Tenenbaum, 2017). Under this proposal, internal models of physical reasoning are necessarily approximate and probabilistic, as they (a) quickly run through and aggregate over the outcomes of many noisy simulations (Smith & Vul, 2013), and (b) take "shortcuts" in order to maintain computational tractability (Vul, Goodman, Griffiths, & Tenenbaum, 2014). The rich mental representations in these models allow for nuanced predictions that flexibly integrate information from variable input sources (Ullman & Tenenbaum, 2020).

Physical Reasoning and Inductive Biases for the Real World at NeurIPS 2021.

The mental simulation account has not gone unchallenged (e.g. Marcus & Davis, 2013), raising questions of the feasibility of adopting such models for AI. Some of this criticism has targeted the notion of an "optimal" or "full" simulation, in which every object in the world is accounted for in the simulation, and its dynamics fully unfolded (Davis & Marcus, 2015). Yet human reasoning is subject to constraints of computation (Lieder & Griffiths, 2020), and mental physical simulation – if it exists – is no exception (Ullman et al., 2017). We suggest that *partial simulation* is one particularly useful approximation, in which humans simulate only objects, events, and properties that they deem relevant. While variations on partial simulation have been proposed (Hegarty, 2004), they have been speculative and not instantiated as computational models. Here we present a specific formal model of such partial simulation, and investigate it qualitatively and quantitatively. We argue such partial simulation is key to efficient implementations of useful commonsense physical reasoning in machines as well.

We take as a case study Ludwin-Peery et al. (2020), which presented particularly compelling empirical evidence challenging the simulation view. They begin with the observation that a full mental simulation should at least roughly mirror basic axioms of probability theory. In contrast to these axioms, Ludwin-Peery et al. (2020) show that people consistently commit a physical version of the *Conjunction Fallacy*: People rate the conjunction of two events $p(A \cap B)$ as *more* likely than its constituent $p(B)$ – a logical impossibility (Tversky & Kahneman, 1982). If physical reasoning involves simultaneous simulation of all the objects in a scene, a conjunction fallacy in the domain of physical reasoning should be impossible; therefore, Ludwin-Peery et al. (2020) argue, physical reasoning cannot rely on simulations.

However, this logic holds only for a full simulation of the scene. We suggest that some people only simulate parts of the scene they are considering. Our model predicts the set of conditions under which one might expect to observe a physical conjunction fallacy. This model rests on only a single parameter, the probability of simulating one of the objects in the scene. Using stimuli and methods that mirror those used by Ludwin-Peery et al. (2020), we collected novel human data to compare against our partial simulation model. We find that our model accurately accounts for people's physical reasoning in a simple prediction task, both qualitatively and quantitatively, including the existence, effect size, and functional form of the physical conjunction fallacy. This model lays the ground for future work exploring how AI systems may also benefit from efficient partial simulation.

2 Model

Under our framework of partial simulation, people only include relevant objects in their mental simulation. Such an implicit decision relies on the pragmatics of the probability judgements people are asked to make. In particular, here we consider probability judgments in the scenarios used by Ludwin-Peery et al. (2020), in which a gray cannonball could collide with a pink sphere (see Fig. 1A). $p(H)$ is the probability of the cannonball Hitting the pink sphere; $p(G)$ is the probability of the pink sphere landing on the Grass; and $p(H \cap G)$ is the conjunction, the probability that the cannonball hits the sphere, and the sphere lands on the grass. Ludwin-Peery et al. (2020) find that $p(H \cap G) > p(G)$ (i.e., a conjunction fallacy). While Ludwin-Peery et al. (2020) did not account for the specific effect size of the conjunction fallacy, we label its magnitude $CF = p(H \cap G) - p(G)$.

Our model crucially involves a term, S, which denotes whether an object (in this case, the cannonball) will be simulated at all. Because the judgment $p(G)$ – *"How likely is it that the pink sphere will end up on the grass?"* – does not explicitly invoke the cannonball, the degree to which the cannonball is simulated (or not) could vary across people and scenes.

If the cannonball is simulated ($S = 1$), then one must consider the cases in which it hits or misses the pink sphere. If the cannonball is not simulated ($S = 0$), one needs only to consider $p(G|\neg H)$). Putting the two options together, and denoting the probability $p(S = 1)$ as $p(S)$ for shorthand, we have:

$$p(G) = p(S)[p(H) * p(G|H) + (1 - p(H)) * p(G|\neg H)] + (1 - p(S)) * p(G|\neg H) \qquad (1)$$

The probability of the cannonball hitting the pink sphere *and* the sphere landing on the grass is:

$$p(H \cap G) = p(H) * p(G|H) \qquad (2)$$

Because simulation of the cannonball is required to calculate $p(H)$, and this is explicitly noted in the phrasing of the question, we assume that the conjunction requires full simulation of both the sphere and the cannonball.

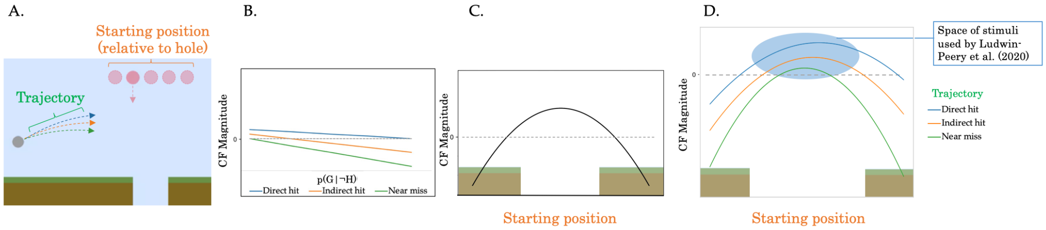

Figure 1: A. Participants saw 15 scenes (3 trajectories × 5 starting positions) in which a cannonball could collide with a pink sphere. B-D. Our model predicts that CF will both increase with more direct-hit trajectories (B), and show an inverse U-shape with the position of the sphere (C), which produces quantitative predictions of the magnitude of the CF effect across all scenes (D).

Our model predicts when and why a conjunction fallacy will be observed. In particular, the magnitude of the conjunction fallacy CF should increase as: (1) $p(H)$ increases (i.e., as the cannonball becomes more likely to directly collide with the pink sphere); (2) $p(G|H)$ increases (i.e., as it becomes more likely that a hit from the cannonball would lead to the pink sphere landing on the grass); and (3) $p(G|\neg H)$ decreases (i.e., as the pink sphere's starting position moves towards the center of the hole). Our model makes the non-obvious predictions that CF will both increase with more direct-hit trajectories, and show an inverse U-shape with the position of the sphere (see Fig. 1D). Notably, the set of stimuli used by Ludwin-Peery et al. (2020) comprised scenes in which the pink sphere was either partially or directly over the hole – circumstances in which a conjunction fallacy would indeed be predicted by our partial simulation model. We test a range of scenes that more fully tiles this space, in order to test a range of our model's predictions.

We use Equations 1 and 2 to build a model that explains the conjunction fallacy in intuitive physics, treating $p(S)$ as a free parameter. We fit this model to aggregate data (over all trials and participants) by performing a grid search over values from 0 to 1, by increments of .01, for $p(S)$. The best-fit values will be those that produce the lowest mean squared error (MSE) between the actual magnitude of conjunction fallacy that participants produced (i.e., $p(H \cap G) - p(G)$), and model predictions for the magnitude of conjunction fallacy, which will be computed using the identities described above based on human ratings of $p(H)$, $p(G|H)$, and $p(G|\neg H)$. Given this best-fit value of $p(S)$, we will examine the relationship between human judgments and model predictions.

3 Methods for Behavioral Study

Participants were adults recruited from Amazon Mechanical Turk via CloudResarch. Participants were paid $4.15, and the study took an average of 23.5 minutes to complete. As stated in our preregistration and mirroring the power analysis in Ludwin-Peery et al. (2020), our final sample consisted of $N = 60$ participants (21 female; $M(SD)_{age} = 38(9.7)$ years). An additional 41 participants were dropped and replaced due to failure to pass built-in check questions.

Using the Pymunk package in Python 3, we created scenes in which a gray cannonball could potentially hit a pink sphere, similar to Ludwin-Peery et al. (2020) in which both objects are above a field of grass with a hole in it. The scenes stop part-way through, after playing for 600 ms – always before the cannonball would hit or miss the pink sphere. These scenes could vary based on two factors: (1) The starting positions of the objects relative to the hole; (2) The trajectory of the gray cannonball. We initially created 54 scenes (9 starting positions × 7 trajectories); from these, we selected a total of 15 videos (5 starting positions × 3 trajectories) to use in our main experiment, based on results from pilot data (see Fig. 1A, and the OSF repository for details).

The general procedure was similar to that used by Ludwin-Peery et al. (2020). After consenting to participate, participants read a detailed description of the task and watched some example videos, to acquaint them with the physical properties of the objects in these scenes. Next, participants viewed all fifteen scenes five times in a blocked design. The order of scenes was randomized across participants. In each of the five blocks, participants were asked to make a different probability judgment for all fifteen scenes: (1) $p(H \cap G)$: "How likely is it that the cannonball will hit the pink sphere, and then the pink sphere will end up on the grass?". (2) $p(G)$: "How likely is it that the pink sphere will end up on the grass?". (3) $p(H)$: "How likely is it that the cannonball will hit the pink sphere?". (4)

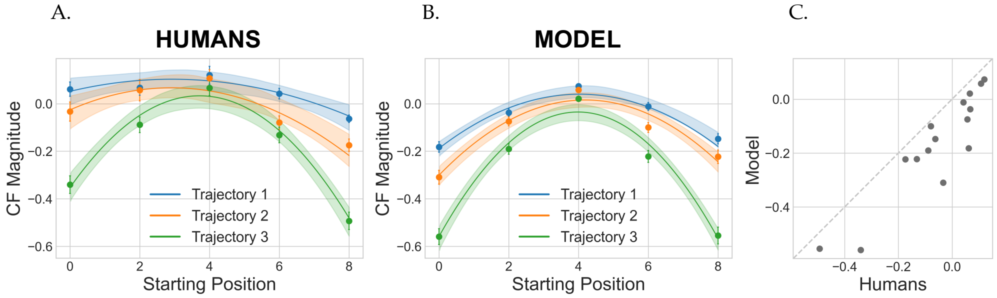

Figure 2: A. Mean human CF values ($p(H \cap G) - p(G)$) by scene, with best-fit polynomials for each trajectory and Bayesian 95% credible interval of the polynomial fit. Error bars = +/- the standard error. B. Mean model predictions for CF values using the best-fit $p(S)$ value $= 0.88$, with best-fit polynomials for each trajectory and Bayesian 95% credible interval of the polynomial fit. Error bars = +/- the standard error. C. The correlation between human data and model predictions was $r(13) = 0.91$.

$p(G|H)$: "Suppose the cannonball hits the pink sphere. How likely is it that the pink sphere would then end up on the grass?". (5) $p(G|\neg H)$: "Imagine that the cannonball was not in the scene at all. How likely is it that the pink sphere would end up on the grass?".

4 Results

We computed CF by first calculating $p(H \cap G) - p(G)$ for each participant on each of the 15 scenes, and then averaging across participants, yielding one aggregated CF value for each scene. First, we performed a one-sample t-test comparing the average CF magnitude to 0, in order to assess whether there was an conjunction fallacy when analyzing our data in the aggregate. The average difference between judgments of $p(H \cap G)$ and $p(G)$ was -0.059, which was significantly different from 0 ($t_{(59)} = -4.57, p < 0.001$). Participants appropriately rated the conjunction as *less* likely than its constituent, thus not succumbing to the conjunction fallacy.

Qualitatively, our model predicts that the magnitude of the conjunction fallacy CF will increase linearly with more direct-hit trajectories, and show an inverse U-shape with position. We first tested these predictions by performing a 2-way repeated measures ANOVA, with the starting position (5 levels) and trajectory (3 levels) of each scene predicting the CF magnitude. Both main effects were significant (Starting position: $F_{(4,216)} = 52.7, p < 0.001, \eta_p^2 = 0.66$; Trajectory: $F_{(2,216)} = 31.5, p < 0.001, \eta_p^2 = 0.54$), as was the interaction ($F_{(8,216)} = 3.4, p = 0.001, \eta_p^2 = 0.11$). Polynomial contrasts were also performed on starting position and trajectory. Supporting our predictions, the best-fit polynomial was quadratic for starting position ($t_{(108)} = -12.8, p < 0.001$), and linear for trajectory ($t_{(54)} = -7.8, p < 0.001$).

The best-fitting value of $p(S)$ was 0.88, which yielded low error (MSE $= 0.017$). We also correlated model predictions with human values for CF, and found excellent fit ($r(13) = 0.91, p < 0.001$; see Fig. 2C). This suggests that the physical conjunction fallacy can be explained by people performing full simulation most of the time, but dropping a relevant object and performing *partial* simulation occasionally.

5 Discussion

Here we suggested that we could better understand failures of probabilistic reasoning for physical events as deriving from *partial* simulation of events, rather than failure to simulate events all together. Our model and the results of our behavioral experiment support this claim and further help to describe the errors in this particular case of physical reasoning. While partial simulation led to errors in judgment in this particular task, we suggest that simulating only "relevant parts" of the scene may contribute to how people perform physical reasoning so efficiently, and thus is beneficial in general for real-world reasoning. As some recent work has begun to demonstrate (Agia et al., 2021; Silver et al., 2021), future developments in AI may consider this form of partial simulation as a potential feature to explore, rather than a bug.

References

Agia, C., Jatavallabhula, K. M., Khodeir, M., Miksik, O., Vineet, V., Mukadam, M., … Shkurti, F. (2021). Taskography: Evaluating robot task planning over large 3D scene graphs. In *Proceedings of5th annual conference on robot learning.*

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*(3), 89-94. doi: 10.1111/j.0963-7214.2004.00281.x

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. doi: 10.1073/pnas.1306572110

Davis, E., & Marcus, G. (2015). The Scope and Limits of Simulation in Cognitive Models. *arXiv preprint arXiv: 1506.04956*, 27.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76. doi: 10.1016/j.cognition.2016.08.012

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

Kominsky, J. F., Strickland, B., Wertz, A. E., Elsner, C., Wynn, K., & Keil, F. C. (2017). Categories and constraints in causal perception. *Psychological Science*, *28*(11), 1649–1662.

Leslie, A. M., Xu, F., Tremoulet, P. D., & Scholl, B. J. (1998). Indexing and the object concept: developing 'what' and 'where' systems. *Trends in cognitive sciences*, *2*(1), 10–18. doi: 10.1016/s1364-6613(97)01113-3

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602-1611. doi: 10.1177/0956797620957610

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, *24*(12), 2351-2360. doi: 10.1177/0956797613495418

Silver, T., Chitnis, R., Curtis, A., Tenenbaum, J., Lozano-Perez, T., & Kaelbling, L. P. (2021). Planning with learned object importance in large problem instances using graph neural networks. In *Proceedings of the 35th AAAI conference on artificial intelligence,*.

Smith, K. A., & Vul, E. (2013). Sources of Uncertainty in Intuitive Physics. *Topics in Cognitive Science*, *5*(1), 185–199. doi: 10.1111/tops.12009

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605–632. doi: 10.1037/0033-295X.99.4.605

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (p. 84–98). Cambridge, England: Cambridge University 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. doi: 10.1016/j.tics.2017.05.012

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533–558. doi: 10.1146/annurev-devpsych-121318-084833

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637. doi: 10.1111/cogs.12101