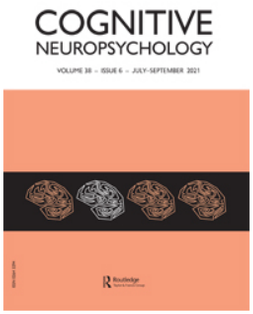

Cognitive Neuropsychology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pcgn20

Partial mental simulation explains fallacies in physical reasoning

Ilona Bass, Kevin A. Smith, Elizabeth Bonawitz & Tomer D. Ullman

To cite this article: Ilona Bass, Kevin A. Smith, Elizabeth Bonawitz & Tomer D. Ullman (2022): Partial mental simulation explains fallacies in physical reasoning, Cognitive Neuropsychology, DOI: 10.1080/02643294.2022.2083950

To link to this article: https://doi.org/10.1080/02643294.2022.2083950

COGNITIVE NEUROPSYCHOLOGY
https://doi.org/10.1080/02643294.2022.2083950

Routledge
Taylor & Francis Group

Partial mental simulation explains fallacies in physical reasoning

Ilona Bass[a,b], Kevin A. Smith[c], Elizabeth Bonawitz[b] and Tomer D. Ullman[a]

[a]Department of Psychology, Harvard University, Cambridge, MA, USA; [b]Graduate School of Education, Harvard University, Cambridge, MA, USA; [c]Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

People can reason intuitively, efficiently, and accurately about everyday physical events. Recent accounts suggest that people use mental simulation to make such intuitive physical judgments. But mental simulation models are computationally expensive; how is physical reasoning relatively accurate, while maintaining computational tractability? We suggest that people make use of *partial simulation*, mentally moving forward in time only parts of the world deemed relevant. We propose a novel partial simulation model, and test it on the *physical conjunction fallacy*, a recently observed phenomenon [Ludwin-Peery et al. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602–1611. https://doi.org/10.1177/0956797620957610] that poses a challenge for full simulation models. We find an excellent fit between our model's predictions and human performance on a set of scenarios that build on and extend those used by Ludwin-Peery et al. [(2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602–1611. https://doi.org/10.1177/0956797620957610], quantitatively and qualitatively accounting for deviations from optimal performance. Our results suggest more generally how we allocate cognitive resources to efficiently represent and simulate physical scenes.

ARTICLE HISTORY
Received 24 November 2021
Revised 8 May 2022
Accepted 25 May 2022

KEYWORDS
Intuitive physics; partial simulation; conjunction fallacy

To interact successfully with the world around us, we need to be able to reason flexibly about how events could unfold – from estimating our chances of success in a round of Pickup Sticks, to gauging the safety of putting just one more dish on top of an already teetering stack, to realizing when a child is about to fall off the balance beam. People implicitly make predictions about the properties of objects (Leslie et al., 1998), and how they will interact (Kominsky et al., 2017), starting as early as infancy (Baillargeon, 2004; Spelke et al., 1992). Yet despite a great deal of theoretical and empirical research, *how* exactly the mind reasons about the unfolding of physical events remains an open question.

Some have suggested that people possess a *mental physics engine*, similar in structure to the programs that run and render physical simulations in modern video games (Battaglia et al., 2013; Hamrick et al., 2016; Téglás et al., 2011). As with many other structured generative cognitive models (Gerstenberg & Tenenbaum, 2017; Tenenbaum et al., 2006, 2011; Ullman & Tenenbaum, 2020), mental game engines support predictions, inference, and generalization. They explain how people are able to make intuitive physical judgments with speed and generality by proposing a mental architecture that uses approximate physical principles to predict what might happen next in an arbitrary scenario (Ullman et al., 2017). Models of mental game engines have the potential to elucidate the developmental trajectory of physical reasoning from infancy to adulthood (Baillargeon, 2004), and to help bridge gaps between human physical judgments and current competencies in artificial intelligence (Lake et al., 2017).

The mental simulation account has not gone unchallenged (e.g., Marcus & Davis, 2013). Earlier theories contended that physical judgments might be better explained by heuristics or rules (Gilden & Proffitt, 1994; Runeson et al., 2000). Later frameworks suggested people intuitively hold pre-Newtonian theories of physics, leading them to consistent misconceptions and biases (e.g., McCloskey et al., 1980,

CONTACT Ilona Bass ibass@fas.harvard.edu Department of Psychology, Harvard University, Cambridge, MA 02138, USA; Graduate School of Education, Harvard University, Cambridge, MA 02138, USA

1983). Many of these biases may be due to modes of presentation, elicited when people are presented with static images or verbal descriptions (Kaiser et al., 1992; Smith et al., 2018). More recently, the notion of a mental game engine has been challenged both empirically and theoretically as being computationally unrealistic, and not accurately accounting for people's inaccuracies in direct intuitive physics tasks (Davis & Marcus, 2015; Ludwin-Peery et al., 2021; Marcus & Davis, 2013).

The apparent conflict between different accounts of intuitive physics may be amiably resolved by recognizing that mental simulations need not be perfect. Previously proposed mental simulation models of physical reasoning are approximate and probabilistic, quickly running through and aggregating over the outcomes of a set of simulations (Battaglia et al., 2013; Smith & Vul, 2013), perhaps only considering a handful at a time (Hamrick et al., 2015; Vul et al., 2014). Beyond these noisy simulations, it has been suggested that mental physics engines take principled short-cuts to maintain computational tractability (Ullman et al., 2017). This explanation is in keeping with the more general argument that human reasoning is subject to constraints of computation (Lieder & Griffiths, 2020); such short-cuts are used in engineered physics engines as well, for similar tractability reasons.

One particularly important general approximation may be that of *partial simulation*. In partial simulation, physical scenes can be handled much more efficiently by only representing and moving forward in time the objects and events that are deemed relevant to the physical judgment at hand. While variations on partial simulation have been proposed in the past (Hegarty, 2004; Ullman et al., 2017), they have not been instantiated as generative computational models. Here, we present a specific formal model of such partial simulation, and investigate it qualitatively and quantitatively. We argue that partial simulation is key to efficient implementations of useful commonsense physical reasoning.

We take as a case study the *physical conjunction fallacy*. A conjunction fallacy occurs when the probability of two events $A \cap B$ happening is judged as being greater than one of the events (A or B) happening on their own (a logical impossibility). People exhibit a conjunction fallacy in a variety of domains, and the phenomenon has attracted a great deal of research and debate (Hertwig & Gigerenzer, 1999; Tversky & Kahneman, 1982). Recently, evidence of the conjunction fallacy was found in judgments of physical events (Ludwin-Peery et al., 2020). If physical reasoning relies on a veridical simulation of physical events, the conjunction fallacy should be impossible. So, by the reasoning of Ludwin-Peery et al. (2020), if the conjunction fallacy is real, physical reasoning cannot be relying on mental simulations. But crucially, this logic holds only for "full" simulation, in which every object in the world is accounted for, and its dynamics fully unfolded.

Here, we propose a partial simulation model of simple two-dimensional scenes (Figure 1). 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. With a single cross-validated parameter (the probability of simulating one of the objects when it is not explicitly cued), we find that our model qualitatively and quantitatively accounts for people's physical reasoning, including the existence, effect size, and functional form of the physical conjunction fallacy. The empirical data collection, analysis, and model comparisons were all pre-registered (see our OSF repository for details).

2. Model

In our framework of partial simulation, people only include *relevant* objects in their mental simulation. Such an implicit decision about what is relevant and what is not relies on the pragmatics of the probability judgements people make. Here, we consider probability judgments involving collisions between simple 2D objects, of the sort used in Ludwin-Peery et al. (2020) (see Figure 1). In these animated scenarios, a gray cannonball is shown en route to hitting a pink sphere that is falling downwards. The pink sphere can end up either falling into a pit, or landing on a green area (the "grass"). People are shown only the first half-second of a scenario unfolding, and are then asked to make various probability judgements about how it will resolve. (For our full set of video stimuli, see the Video Stimuli sub-folder of our OSF repository.)

We label $p(H)$ the probability of the cannonball Hitting the pink sphere, and $p(G)$ the probability of the pink sphere landing on the Grass. The probability

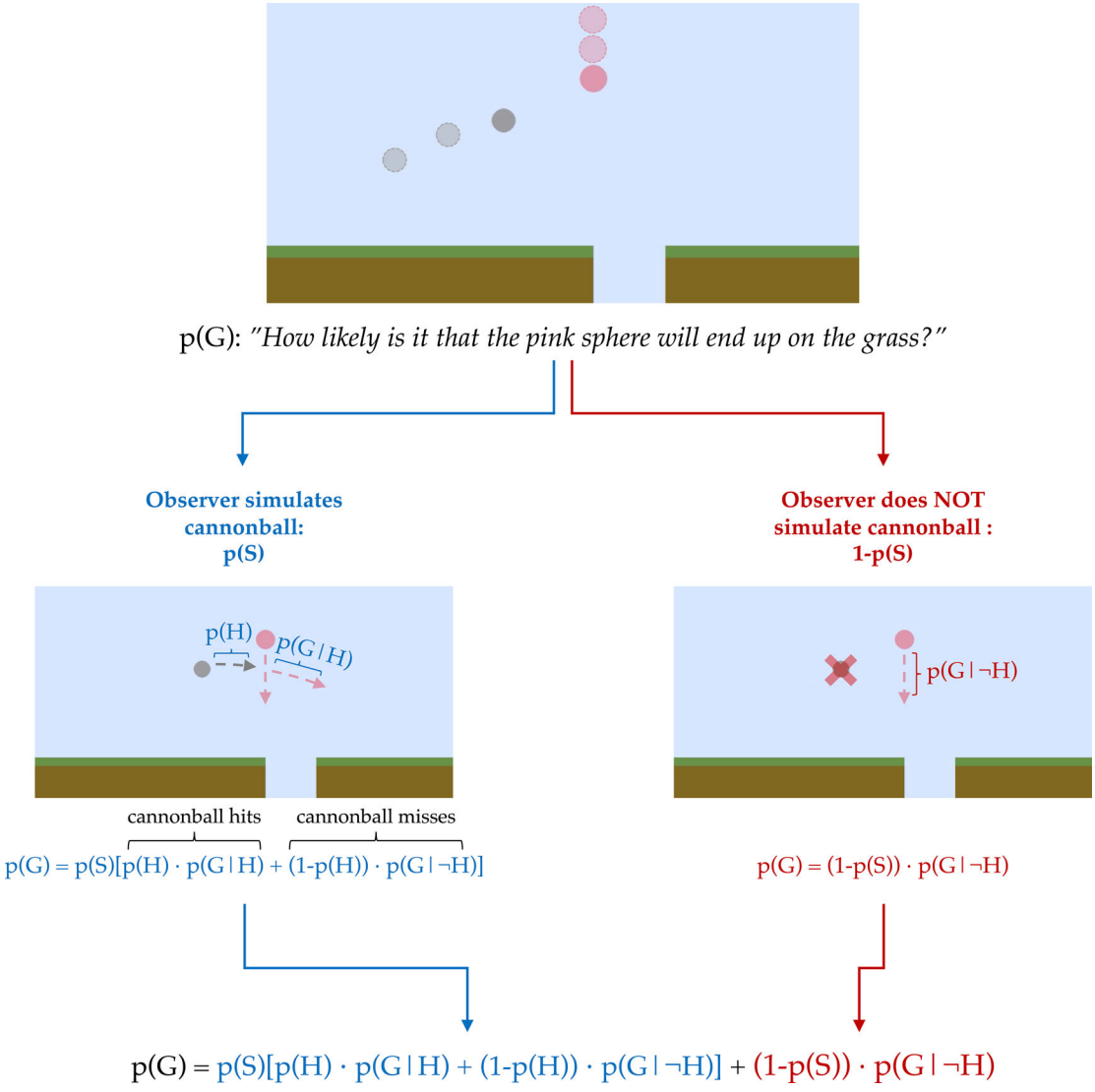

Figure 1. Model sketch using snapshots of stimuli. In our scenarios, a gray cannonball moves towards a pink sphere that is falling downwards. When asked to judge a situation that does not explicitly invoke an object, people may or may not simulate that object. Specifically, when asked to judge $p(G)$ *How likely is it that the pink sphere will end up on the grass?*, people may simulate the motion of the cannonball ($S = 1$) or not ($S = 0$), with some probability, $p(S)$. If people simulate the cannonball (left), they must consider the case in which the cannonball hits the pink sphere, and the case in which it misses. When not simulating (right), people need to reason about only one situation, equivalent to the case of the cannonball missing. Together, these possibilities lead to Equation (1), the predicted probability judgement of the pink sphere ending up on the grass. If $p(S) < 1$, $p(G)$ deviates from an optimal simulation, possibly leading to a conjunction fallacy. [To view this figure in colour, please see the online version of this journal.]

of both of these events occurring (the cannonball hits the sphere, and the sphere lands on the grass) is the conjunction $p(H \cap G)$. Ludwin-Peery et al. (2020) found a physical conjunction fallacy, specifically that people judge $p(H \cap G)$ to be greater than $p(G)$. We label the magnitude of this *probability difference* as $P_{dif} = p(H \cap G) - p(G)$. When $P_{dif} > 0$, a conjunction fallacy will be observed.

Crucially, our model also involves a term, S, which denotes whether an object (in this case, the cannonball) will be simulated at all. Judgments that directly involve an object (e.g., the cannonball in the judgment, *How likely is it that the cannonball will hit the pink sphere?*) necessarily set $S = 1$ for that object. However, judgements that do not directly involve an object may have $S = 0$. We denote the probability that an object is simulated as $p(S = 1)$, or $p(S)$ for short. 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 probability

that the cannonball is simulated (or not) could vary across different people, probability judgments, and scenes.

If a person simulates the motion of the cannonball ($S = 1$), then they must consider the cases in which it hits or misses the pink sphere. If the cannonball is not simulated ($S = 0$), the situation is equivalent to the cannonball missing the sphere, $p(G \mid \neg H)$. See also Figure 1. Putting the two options together we have:

$$p(G) = p(S) \cdot [p(H) \cdot p(G \mid H) + (1 - p(H)) \cdot p(G \mid \neg H)] + (1 - p(S)) \cdot p(G \mid \neg H). \quad (1)$$

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

$$p(H \cap G) = p(G \mid H) \cdot p(H). \quad (2)$$

Because simulating the cannonball is required to calculate $p(H)$, and the cannonball is explicitly noted in the phrasing of the question, we assume that the conjunction requires full simulation of both the sphere and the cannonball ($S = 1$).

By expressing $p(G)$ as Equation (1) and $p(H \cap G)$ as Equation (2), our model can predict when, why, and to what degree a conjunction fallacy will be observed (where $P_{dif} = p(H \cap G) - p(G)$, and a conjunction fallacy occurs when $P_{dif} > 0$). Recall that neglecting to simulate the cannonball (i.e., $S = 0$) effectively sets $p(H) = 0$. Therefore, when the cannonball is on a collision course with the pink sphere, participants should underestimate $p(G)$ relative to $p(H \cap G)$, leading to an increase in P_{dif}. When a hit from the cannonball would make it likely for the pink sphere to end up on the grass – as when the pink sphere starts either over the hole or over the grass on the side further from the cannonball – failing to simulate the cannonball will again lead to an increase in P_{dif}, for similar reasons. And when the pink sphere's starting position is such that no hit from the cannonball (i.e., a straight-down drop) would result in it landing in the grass, differences between judgments that result from different degrees of simulation should become compressed, thereby decreasing P_{dif}. So, our model makes the non-obvious predictions that P_{dif} will increase with more direct-hit trajectories (i.e., as $p(H)$ increases; see Figure 2(B,C)), and decrease as the pink sphere's starting position moves further away from the centre of the hole (i.e., as $p(G \mid \neg H)$ increases; see Figure 2(D)). Together, this results in an inverse U-shape with the position of the sphere that is modulated by the likelihood of a collision (see Figure 2(E)).

What's more, Equation (1) allows us to make intermediate predictions about $p(G)$ in particular, providing a more detailed quantification of possible drivers of the physical conjunction fallacy. Here, our model makes asymmetric predictions: When the pink sphere is over the grass to the side *further* from the cannonball, we expect a slight *under*-prediction of $p(G)$, because any collisions will cause it to be more likely to land on the grass – but overall, the probability of the pink sphere landing on the grass is high regardless of whether or not a hit from the cannonball was simulated. On the other hand, when the pink sphere is over the grass on the side *closer* to the cannonball, we expect a relative *over*-prediction of $p(G)$, because any failure to simulate the cannonball makes it more likely that the pink sphere will drop straight down onto the grass rather than getting knocked into the hole.

These qualitative trends are not noted in or explained by the previous findings (Ludwin-Peery et al., 2020). Importantly, the set of stimuli used by Ludwin-Peery et al. (2020) was made up of scenes in which the pink sphere was either partially or directly over the hole – circumstances in which a conjunction fallacy is most strongly predicted by our partial simulation model. Here, we use 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 partial simulation model, 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 of $p(S)$ from 0 to 1, by increments of 0.01. The best-fit values are those that produce the lowest root mean squared error (RMSE) between the actual probability difference that participants produced (i.e., $P_{dif} = p(H \cap G) - p(G)$), and model predictions for this probability difference, which are computed using the identities described above (Equations (1) and (2)) based on human ratings of $p(H)$, $p(G \mid H)$, and $p(G \mid \neg H)$. As we discuss in the Method section below, we assess people's judgments of each component probability ($p(H \cap G)$, $p(G)$, $p(H)$, $p(G \mid H)$, and $p(G \mid \neg H)$) separately on distinct trials. Thus, we can investigate the extent to which these identities hold with a single free parameter, and whether this best-

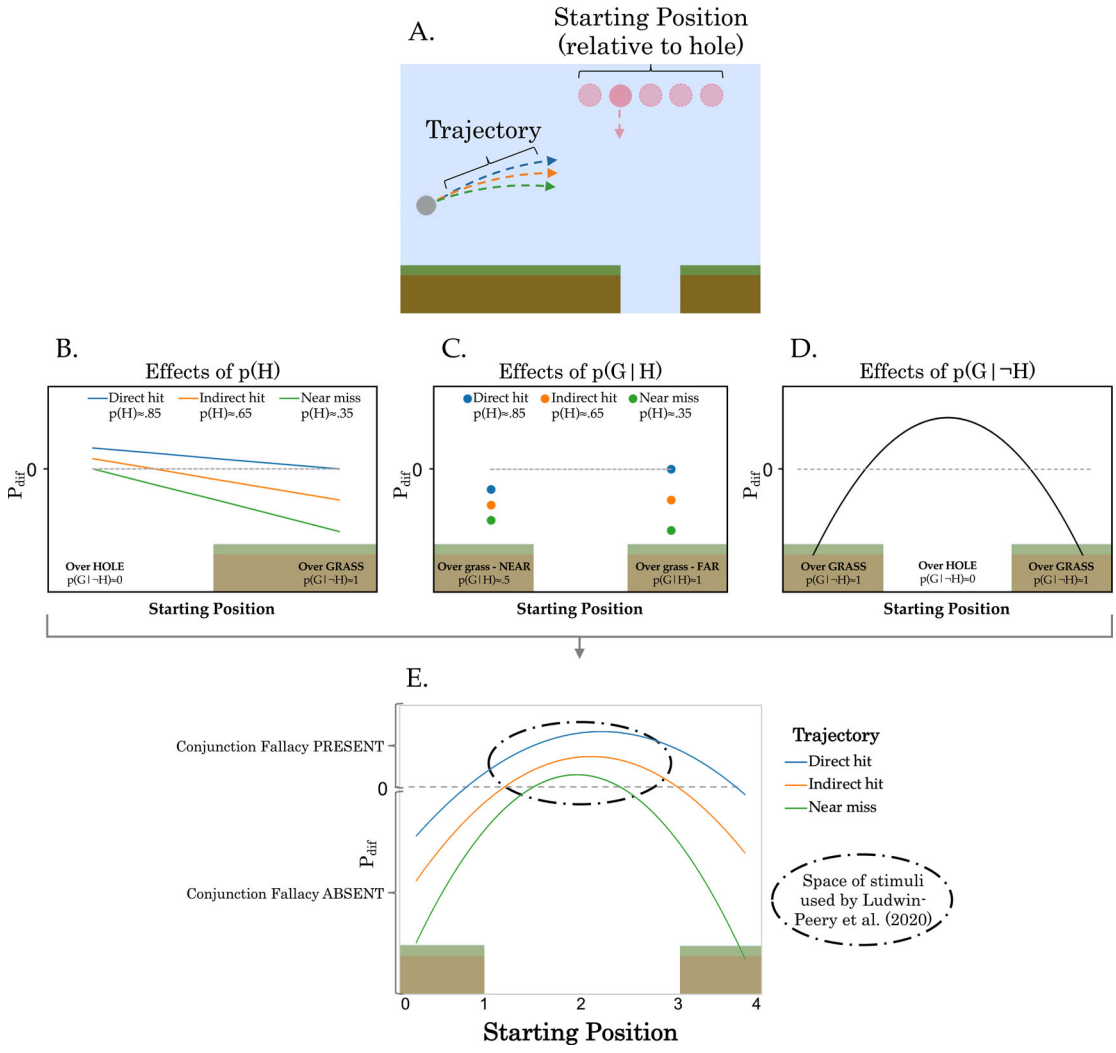

Figure 2. Stimuli sketch and model predictions. (A) Participants saw 15 scenes (3 trajectories × 5 starting positions) in which a cannonball could collide with a pink sphere. (B) Our model predicts that P_{dif} (i.e., $p(H \cap G) - p(G)$) will increase with more direct-hit trajectories, (C) be more compressed when the pink sphere starts over the grass on the side closer to the cannonball, and (D) show an inverse U-shape with the position of the sphere. (E) Combining these three produces quantitative predictions of the magnitude of the conjunction fallacy effect (or lack thereof) across all scenes. We have also highlighted the approximate space of stimuli used by Ludwin-Peery et al. (2020). Predictions and zero-level of P_{dif} were calculated using the $p(S)$ value fit to data as described below, but qualitative trends are invariant to this parameter setting. [To view this figure in colour, please see the online version of this journal.]

fit value of $p(S)$ allows the model to make accurate predictions of human behaviour.

3. Method

3.1. Participants

Participants were recruited from Amazon Mechanical Turk via CloudResearch. Participants were paid $4.15, and the study took an average of 23.5 min 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; see Procedure below for details. (This drop rate is comparable to that in Ludwin-Peery et al., 2020, and is not atypical for Mechanical Turk studies; see Zhou & Fishbach, 2016.)

3.2. Materials

Using the Pymunk API for the Chipmunk 2D physics engine (Lembcke, 2013), we created short animations in which a gray cannonball could potentially hit a pink

sphere, similar to Ludwin-Peery et al. (2020). In these scenes, both objects are above a field of grass with a hole in it (see Figure 1). The animations played for 600 ms, and stopped before the cannonball would hit or miss the pink sphere. The scenes varied by (1) The starting positions of the objects relative to the hole, and (2) The trajectory of the gray cannonball. We created a total of 15 videos (5 starting positions × 3 trajectories) to use in our main experiment, based on results from pilot data (see Figure 2(A), and our OSF repository for details). The starting position was determined by the pink sphere's initial location relative to the hole. The starting position of the gray cannonball relative to the pink sphere was identical across scenes. The five starting locations varied by increments of 60 pixels, such that the pink sphere started either (1) over the grass to the left[1] of the hole, (2) over the left corner between the grass and the hole, (3) directly over the hole, (4) over the right corner between the grass and the hole, or (5) over the grass to the right of the hole. The trajectory of the gray cannonball was determined by how it would hit or miss the pink sphere: (1) a direct hit, (2) an indirect hit, or (3) a near miss.[2]

3.3. Procedure

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. Participants then answered nine simple comprehension questions about the task, which were very similar to those used by Ludwin-Peery et al. (2020). (See our OSF repository for details about these comprehension checks.) Participants who were unable to answer all nine of these questions correctly after three attempts were excluded from subsequent analysis and replaced ($N = 19$).

Next, participants viewed all fifteen scenes five times in a blocked design. The order of scenes viewed within each block was randomized across participants, and half of the videos in each block were mirrored horizontally. 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) $p(G \mid H)$: "Suppose the cannonball hits the pink sphere. How likely is it that the pink sphere would then end up on the grass?"
(5) $p(G \mid \neg H)$: "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?"

For each probability judgment, participants first played the video clip – which they had the option of watching as many times as they needed – and then provided their rating by dragging a slider on a scale from 0 to 100 (starting position of 50). The last frame of the clip remained on screen when participants were making their judgments. (This is identical to the procedure used by Ludwin-Peery et al., 2020.)

To ensure that the two judgments that contribute to the conjunction fallacy metric were not biased by other judgments, the $p(H \cap G)$ and $p(G)$ blocks were presented first (order counterbalanced), followed by the $p(H)$, $p(G \mid H)$, and $p(G \mid \neg H)$ blocks (order counterbalanced).

After completing the main experiment (which consisted of a total of $15 \times 5 = 75$ probability judgments), participants provided demographic information, and had the opportunity to give qualitative feedback about the task. In this section, we asked participants if they experienced any technical difficulties. Participants were then debriefed, and thanked for participating.

We used (pre-registered) exclusion criteria to drop and replace participants who: failed to answer comprehension questions (as mentioned earlier, $N = 19$); indicated that they were not able to see every video presented to them throughout the experiment ($N = 8$); or rated $p(G \mid \neg H)$ (*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?*) as higher than 25% for any of the three scenes in which the pink sphere started directly over the hole ($N = 14$). Also, prior to analysing data in the aggregate, we dropped individual data points that were more than two standard deviations away from the mean for that rating on a particular scene.

4. Results

We computed the magnitude of the probability difference by first calculating $p(H \cap G) - p(G)$ for each participant on each of the 15 scenes, and then averaging across participants. This gave one aggregated P_{dif} value for each scene.

4.1. Model-free analyses

First, we performed a one-sample t-test comparing the average P_{dif} magnitude to 0, in order to assess whether there was a conjunction fallacy when analysing our data in the aggregate. The average difference between judgments of $p(H \cap G)$ and $p(G)$ was −0.059, which was significantly less than 0 ($t_{(59)} = -3.91$, $p < 0.001$, two-tailed). Over all of the scenarios tested here, participants did not show a conjunction fallacy; instead, participants appropriately rated the conjunction as *less* likely than its constituent. This does *not* conflict with the findings of Ludwin-Peery et al. (2020), as we presented participants with a mixture of trials where we would and would not expect the conjunction fallacy to appear. Indeed, in the subset of trials that roughly maps on to the stimuli set used by Ludwin-Peery et al. (2020) (all three trajectories in starting position 2, and the direct-hit trajectory in starting position 3; see Figure 2(E)), we do find an overall conjunction fallacy ($P_{dif} = 0.090$; $t_{(59)} = 4.19$, $p < 0.001$, two-tailed).

Qualitatively, our model predicts that the magnitude of the probability difference $P_{dif} = p(H \cap G) - p(G)$ will increase linearly with more direct-hit trajectories, and show an inverse U-shape with position. As can be seen in (Figure 3(A)), both of these predictions were confirmed. We statistically 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 P_{dif}. 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 ($t_{(54)} = 7.8$, $p < 0.001$).

4.2. Model-based analyses

As described in the Model section above, we treated $p(S)$ as a free parameter and found the value that produced the lowest RMSE between participant-produced P_{dif} and model-predicted P_{dif}. The best-fitting value of $p(S)$ was 0.87, which yielded low error (RMSE = 0.13). Correlations between model predictions and human values for P_{dif} showed an excellent fit ($r_{(13)} = 0.91$, $p<0.001$; see Figure 3(A–C)). We found good reliability of this best-fit value for $p(S)$ in 1000 bootstrapped samples of parameter estimates (95% CI = [0.79, 0.95]). We also looked separately at our model's ability to predict participant-produced $p(G)$, given the same best-fit value of $p(S)$ found above. Again, we found an outstanding fit between model predictions and human performance (RMSE = 0.069; $r_{(13)} = 0.94$, $p<0.001$; see Figure 3(D–F)).

Finally, we investigated our model's ability to predict participant-P_{dif} at the individual level, both by participant and by scene. To do this, we computed RMSE between model predictions for P_{dif} and participant-produced P_{dif} using cross-validated values for $p(S)$, in three different ways: (1) Aggregating over all participants, fitting a single $p(S)$ to half of the scenes; (2) Allowing $p(S)$ to vary by participant, finding best-fit $p(S)$ values for each participant on half of the scenes; and (3) Allowing $p(S)$ to vary by scene, finding best-fit $p(S)$ values on each scene for half of the participants. We then compared aggregate RMSE to participant RMSE and trial RMSE, in order to assess the extent to which our model could explain individual differences in P_{dif} (where ΔRMSE = individual RMSE – aggregate RMSE). We found that individual participant fits only outperformed aggregate fits on 62% of the 1000 cross-validation runs (average ΔRMSE = 0.163; 95% CI = [−0.384, 0.999]); and trial fits only outperformed aggregate fits 8% of the time (average ΔRMSE = −0.072; 95% CI = [−0.116, −0.014]). These results can come about for three reasons, which are not mutually exclusive. First, model fits on the aggregate data were already near ceiling, making it difficult to detect whether a model fit to individual data improves the fit. Second, individuals may indeed vary in their likelihood of simulating the full scene trial by trial. Third, our design required relatively few trials per participant, leading to limited data for the purposes of

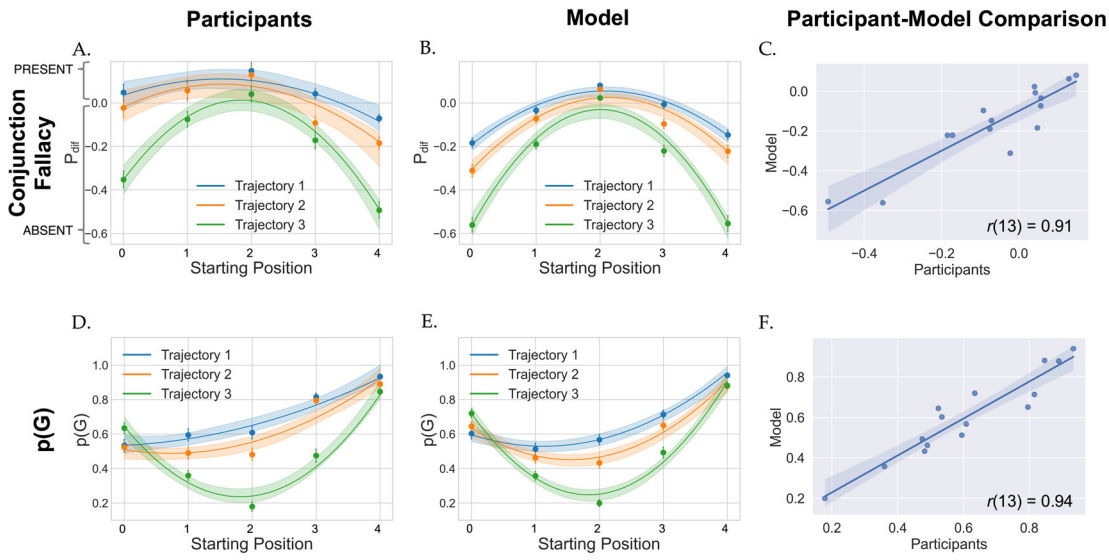

Figure 3. Empirical results and comparison to model. (A) Mean participant P_{dif} values ($p(H \cap G) - p(G)$) by scene, with best-fit polynomials for each trajectory and Bayesian 95% credible interval of the polynomial fit. Error bars $= +/-$ the standard error. When $P_{dif} > 0$, a conjunction fallacy is present; when $P_{dif} < 0$, it is absent. (B) Mean model predictions for P_{dif} values using the best-fit $p(S)$ value $= 0.87$, 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 participant data and model predictions for P_{dif}, with bootstrapped 95% confidence interval around the best-fit line ($r(13) = 0.91$). (D) Mean participant $p(G)$ values by scene, with best-fit polynomials for each trajectory and Bayesian 95% credible interval of the polynomial fit. Error bars $= +/-$ the standard error. (E) Mean model predictions for $p(G)$ values using the best-fit $p(S)$ value $= 0.87$, with best-fit polynomials for each trajectory, and Bayesian 95% credible interval of the polynomial fit. Error bars $= +/-$ the standard error. (F) The correlation between participant data and model predictions for $p(G)$, with bootstrapped 95% confidence interval around the best-fit line ($r(13) = 0.94$). [To view this figure in colour, please see the online version of this journal.]

individual model fits. Future work can explore the role of individual differences or scene effects with more trials.

Taken together, our results suggest that the physical conjunction fallacy can be qualitatively and quantitatively explained by most people performing full simulation most of the time, and occasionally performing partial simulation by not moving an uncued object forward in time.[3]

5. Discussion

Intuitive physical reasoning is a fundamental part of everyday life, from mundane activities, to flexible responses in novel situations, to planning complex actions. The mental physics engine framework proposes that the accuracy and efficiency of intuitive physics can be explained by an approximate simulation of possible outcomes. Counter-proposals argue that mental physics engines are theoretically and empirically dubious. Recent evidence found that physical reasoning is subject to logical fallacies that violate the axioms of probability. If mental physics engines are polished mirrors of reality, such fallacies are impossible. But we see through a glass, darkly. Mental engines do not have to be perfect. Like anything human and taxed by computational resource limitations, they almost certainly are not.

Here, we focussed on a particularly useful approximation that may be operating in mental physics engines – *partial simulation*. We took as a case study the physical conjunction fallacy, and found that it could be well explained by some people failing to simulate part of a scene. We presented a model that instantiated this theory, and found a good correlation between our model's predictions and people's responses. Beyond simple correlations, our model uses a single variable to explain several empirical findings: why the conjunction fallacy is of the particular magnitude that it is; why it presents for some scenarios and not others; why it depends linearly on the trajectory of one object, and quadratically on initial position of a second object. Such empirical findings can seem disparate and puzzling, but turn out to come naturally from a single, simple, partial simulation. Of course, this work is not the first to

propose partial simulation as a cognitive strategy in intuitive physics (Hegarty, 2004; Ullman et al., 2017). However, our formalization of this phenomenon is the first that quantitatively explains how the effects of partial simulation could be brought to bear in everyday reasoning.

It is important to emphasize that we are not contesting the findings from Ludwin-Peery et al. (2020) – quite the opposite, as their results reflect exactly what we would expect to see under our partial simulation account. We are encouraged by this scientific discourse, which exemplifies the tenets of open research. Ludwin-Peery et al. (2020) raised a reasonable objection to mental simulation theories, and presented compelling empirical evidence to support their objection. They made the deliberate decision to make their study materials and data openly available through online repositories, which helped us immensely in developing our model and experiments. Ultimately, the synthesis of this cycle (proposals leading to challenges leading to revised proposals) is precisely the process that is going to help drive forward the development of our computational models and theories of cognition. We are excited for the scientific understanding of intuitive physics that we expect to emerge from these debates: A model of physical reasoning that includes "shortcuts" the mind might take to overcome cognitive limitations, that can in turn explain the set of errors and biases that we display when reasoning about physical events.

Our partial simulation model provides insight into the cognitive processes that underlie physical reasoning abilities, and explains possible sources of error in these judgments. But, we are *not* attempting to directly explain the origins of the conjunction fallacy per se. While we use the physical conjunction fallacy as a case study here, the conjunction fallacy is a pervasive phenomenon – and many of its instantiations are unlikely to require mental simulation. However, we do believe our model and others like it may be able to help account for patterns in other kinds of physical judgments. This includes the potential to explain deviations from perfect prediction found in developmental studies, such as when young children fail to take the presence of a barrier into account when making predictions regarding the final trajectory of a moving object (e.g., Hood et al., 2000).

In this paper, we used one type of scenario from Ludwin-Peery et al. (2020) in which a cannonball could collide with a pink sphere. But in fact, Ludwin-Peery et al. (2020) found evidence of a physical conjunction fallacy across other kinds of scenes, as well (e.g., block towers; a ball bouncing off a backboard and into a bucket; etc.). And other forms of "partial simulation" might explain more classical errors of physical reasoning – e.g., the fact that people believe an object dropped from an airplane will fall straight down (McCloskey et al., 1983), which has traditionally been considered an error in the reference frame of the plane, might instead be explained as a simulation that ignores horizontal velocity. Testing the extent to which a more general form of our model applies across a variety of scenarios could help us understand the ubiquity of partial simulation in physical reasoning more broadly.

While our model provided excellent fit to our empirical data, there are also interesting points of divergence. For instance, although model predictions and participant responses both reflected a similar qualitative asymmetry in the compression of P_{dif} between trajectories for starting positions over the grass on the near versus far side of the hole, our model predicts that there should be little to no conjunction fallacy for starting positions 0 and 1 (see Figure 3(B)). Yet in our empirical data, we see that in the aggregate, participants *did* in fact commit a conjunction fallacy in these cases (see Figure 3(A)). It is possible – even likely – that cognitive processes other than partial simulation could explain errors in physical reasoning. For example, lower-level perceptual features of scenes (e.g., the salience of particular objects relative to others) may influence attention and subsequent memory for starting positions and trajectories, which could systematically bias probability judgments. It could also be that perceptual cues contribute to implicit decisions about whether or not to simulate. Digging deeper into the cases in which our model diverges from people's judgments is a fruitful avenue for future research.

Our findings have promising implications for AI models of physical reasoning. A flexible system for reasoning about physical events is a fundamental aspect of desired machine intelligence (Kuipers, 1986). Yet our current state-of-the-art engineered systems come nowhere near human competence on these tasks. Recent algorithms have demonstrated

circumscribed success in modelling specific situations like how a stack of objects will fall (Groth et al., 2018) or how fluids will pour (Sanchez-Gonzalez et al., 2020). But these same models often fail to generalize outside of the scenarios they were trained on Bear et al. (2021), and it is a challenge to get them to make useful predictions in the real world (Kloss et al., 2020). These systems may be hitting a particular stumbling block, in that they are engineered to make veridical predictions of *all* aspects of a dynamic scene, often to the point of predicting images at the pixel level (Babaeizadeh et al., 2021). By contrast, efficient simulation means choosing the right idealization (Davis & Marcus, 2015; Fisher, 2006). In practice, systems that predict and plan over appropriately reduced representations are also more efficient (Agia et al., 2021; Silver et al., 2021). AI systems that reason more reasonably about physics could benefit from incorporating the same shortcuts that humans might be using, including limited samples (Battaglia et al., 2013; Hamrick et al., 2015), simplified shape representations (Smith et al., 2019; Ullman et al., 2017), or partial simulation as studied here.

In line with recent work on grounded commonsense reasoning (Bisk et al., 2020; Yi et al., 2019), the current findings highlight the importance of building formal frameworks that take seriously the combination of language, pragmatics, and physics. Under our model of partial simulation, people might not simulate an object when that object is not invoked by pragmatics. It is striking that, despite the prominent role of the cannonball in our scenes, participants sometimes deemed it irrelevant enough to the probability judgment at hand that they did not simulate it at all. This suggests that people may be making rich inferences from the pragmatics of the question itself, drastically constraining the space of simulations that could thus result. Of course, when and why people might fully versus partially simulate a scene remains an open question. Partial simulation likely relies on some interplay between particular scene features, and pragmatic inferences. To really understand the role of language in implicit simulation decisions and physical judgments more broadly, we will need models that translate the pragmatics of the question into the relevant simulation. This has long been a challenge for mental physics engines: When asked whether a tower will fall, what exactly do people think is meant by "fall"? (See Battaglia et al., 2013.) Even in the absence of linguistic cues, people may use implicit practical considerations of what is relevant enough to mentally move forward. How considerations of relevance are made are, however, outside the scope of this paper. Deviations from optimal physical reasoning can come about from a *partial simulation*, rather than a failure to simulate things at all. Our model and behavioural findings support this claim, and help explain a specific pattern of deviations in a well-described physical reasoning situation. If we had world enough, and time, full simulation were no crime. Partial simulation can lead to errors in judgment, but on the whole can get things reasonably right, reasonably fast. Much like concluding sentences, simulations don't have to be perfect; they can just be good enough.

Notes

1. Across all of the scenes we created, the cannonball starts over the grass to the left side of the hole, and its trajectory is left-to-right. Therefore, the grass on the left side of the hole is the "near" side with respect to the cannonball, and the grass on the right side of the hole is the "far" side. As we describe in the Procedure below, we also mirrored half of the trials horizontally; in these cases, the right side of the hole was closer to the cannonball.
2. Separately collected pilot data certainty ratings tie these hit/miss characterizations directly to psychological certainty judgements.
3. Because we had limited data for the purposes of individual model fits, we cannot say the degree to which $p(S) = 0.87$ represents 87% of participants simulating across all trials, or 87% of the trials being fully simulated by all participants, or some mixture of the two. Rather, we can only infer that a partial simulation model in which full simulation occurs at the group level 87% of the time provides the best quantitative fit to our empirical data.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

IB is supported by a Mind Brain Behavior postdoctoral fellowship from Harvard University. Thanks to the James S. McDonnell Foundation Scholar Award in Understanding Human Cognition (EB). TDU and KAS are supported by NSF Science Technology Center Award CCF-1231216 and the

Defense Advanced Research Projects Agency [DARPA] Machine Common Sense program. KAS is supported by NSF grant: Adversarial Collaborative Research on Intuitive Physical Reasoning (#2121009).

Funding

IB is supported by a Mind Brain Behavior postdoctoral fellowship from Harvard University. Thanks to the James S. McDonnell Foundation Scholar Award in Understanding Human Cognition (EB). TDU and KAS are supported by NSF Science Technology Center Award CCF-1231216 and the Defense Advanced Research Projects Agency [DARPA] Machine Common Sense program. KAS is supported by NSF grant: Adversarial Collaborative Research on Intuitive Physical Reasoning (#2121009).

References

Agia, C., Jatavallabhula, K. M., Khodeir, M., Miksik, O., Vineet, V., Mukadam, M., Paull, L., & Shkurti, F. (2021). Taskography: Evaluating robot task planning over large 3D scene graphs. In A. Faust, D. Hsu, & G. Neumann (Eds.), *Proceedings of the 5th conference on robot learning* (pp. 46–58). PMLR.

Babaeizadeh, M., Saffar, M. T., Nair, S., Levine, S., Finn, C., & Erhan, D. (2021). *FitVid: Overfitting in pixel-level video prediction*. Preprint. arXiv:2106.13195.

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*(3), 89–94. https://doi.org/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. https://doi.org/10.1073/pnas.1306572110

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H. Y. F., R. T. Pramod, Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., Fei-Fei, L., Kanwisher, N., Tenenbaum, J. B., Yamins, D. L. K., & Fan, J. E. (2021). *Physion: Evaluating physical prediction from vision in humans and machines*. Preprint. arXiv:2106.08261.

Bisk, Y., Zellers, R., Gao, J., & Choi, Y. (2020). PIQA: Reasoning about physical commonsense in natural language. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, pp. 7432–7439). AAAI Press.

Davis, E., & Marcus, G. (2015). *The scope and limits of simulation in cognitive models*. Preprint. arXiv: 1506.04956.

Fisher, J. C. (2006). Does simulation theory really involve simulation? *Philosophical Psychology*, *19*(4), 417–432. https://doi.org/10.1080/09515080600726377

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

Gilden, D. L., & Proffitt, D. R. (1994). Heuristic judgment of mass ratio in two-body collisions. *Perception & Psychophysics*, *56* (6), 708–720. https://doi.org/10.3758/BF03208364

Groth, O., Fuchs, F. B., Posner, I., & Vedaldi, A. (2018). ShapeStacks: Learning vision-based physical intuition for generalised object stacking. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *ECCV 2018* (Vol. 11205, pp. 724–739). Springer International Publishing. https://doi.org/10.1007/978-3-030-01246-5_43

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*(45), 61–76. https://doi.org/10.1016/j.cognition.2016.08.012.

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, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 866–871). Austin, TX: Cognitive Science Society.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285. https://doi.org/10.1016/j.tics.2004.04.001

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*(4), 275–305. https://doi.org/10.1002/(SICI)1099-0771(199912)12:4¡275::AID-BDM323¿3.0.CO;2-M

Hood, B., Carey, S., & Prasada, S. (2000). Predicting the outcomes of physical events: Two-year-olds fail to reveal knowledge of solidity and support. *Child Development*, *71*(6), 1540–1554. https://doi.org/10.1111/cdev.2000.71.issue-6

Kaiser, M. K., Proffitt, D. R., Whelan, S. M., & Hecht, H. (1992). Influence of animation on dynamical judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 669–689. https://doi.org/10.1037//0096-1523.18.3.669

Kloss, A., Schaal, S., & Bohg, J. (2020). Combining learned and analytical models for predicting action effects from sensory data. *The International Journal of Robotics Research*, *00*(0), 1–20. https://doi.org/10.1177/0278364920954896.

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. https://doi.org/10.1177/0956797617719930

Kuipers, B. (1986). Qualitative simulation. *Artificial Intelligence*, *29*(3), 289–338. https://doi.org/10.1016/0004-3702(86)90073-1

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*(E253), 1–72. https://doi.org/10.1017/S0140525X16001837.

Lembcke, S. (2013). *Chipmunk 2D physics engine*. Howling Moon Software.

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. https://doi.org/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*(E1), 1–60. https://doi.org/10.1017/S0140525X1900061X.

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. https://doi.org/10.1177/0956797620957610

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). *Cognitive Psychology*, *127*(1), 101396. https://doi.org/10.1016/j.cogpsych.2021.101396.

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

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. *Science*, *210*(4474), 1139–1141. https://doi.org/10.1126/science.210.4474.1139

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 636–649. https://doi.org/10.1037/0278-7393.9.4.636

Runeson, S., Juslin, P., & Olsson, H. (2000). Visual perception of dynamic properties: Cue heuristics versus direct-perceptual competence. *Psychological Review*, *107*(3), 525–555. https://doi.org/10.1037/0033-295x.107.3.525

Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., & Battaglia, P. (2020). Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 8459–8468). PMLR.

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 AAAI conference on artificial intelligence* (Vol. 35, pp. 11962–11971). AAAI Press.

Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different physical intuitions exist between tasks, not domains. *Computational Brain & Behavior*, *1*(2), 101–118. https://doi.org/10.1007/s42113-018-0007-3

Smith, K. A., Mei, L., Yao, S., Wu, J., Spelke, E. S., Tenenbaum, J. B., & Ullman, T. D. (2019). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 8985–8995). Curran Associates, Inc.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199. https://doi.org/10.1111/tops.12009

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

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–1059. https://doi.org/10.1126/science.1196404

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). 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. https://doi.org/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*(1), 533–558. https://doi.org/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. https://doi.org/10.1111/cogs.12101

Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., & Tenenbaum, J. B. (2019). *CLEVRER: Collision events for video representation and reasoning*. Preprint. arXiv:1910.01442.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056