Limits on the Use of Simulation in Physical Reasoning

Ethan Ludwin-Peery[1] (elp327@nyu.edu), **Neil R. Bramley**[2] (Neil.Bramley@ed.ac.uk)
Ernest Davis[3] (davise@cs.nyu.edu), **Todd M. Gureckis**[1] (todd.gureckis@nyu.edu)
[1]Department of Psychology, NYU, New York, [2]Department of Psychology, University of Edinburgh, Edinburgh, Scotland,
[3]Department of Computer Science, NYU, New York

Abstract

In this paper, we describe three experiments involving simple physical judgments and predictions, and argue their results are generally inconsistent with three core commitments of probabilistic mental simulation theory (PMST). The first experiment shows that people routinely fail to track the spatio-temporal identity of objects. The second experiment shows that people often incorrectly reverse the order of consequential physical events when making physical predictions. Finally, we demonstrate a physical version of the conjunction fallacy where participants rate the probability of two joint events as more likely to occur than a constituent event of that set. These results highlight the limitations or boundary conditions of simulation theory.

Keywords: intuitive physics; mental simulation; inference; conjunction fallacy

Introduction

Successful interaction with our environment often requires reasoning about the physical world (e.g., predicting if a stack of books on a desk is unstable), but the mental processes that support this ability remain poorly understood. Simulation is a technique used for physical reasoning in many applications ranging from modeling molecular interactions to designing realistic video game physics engines. In a simulation, a program starts with an initial state, and then applies the relevant dynamic laws of physics to compute what will happen over a series of (typically short) time steps; in effect, computing a "movie" of how the scenario progresses.

Some researchers have recently argued that humans use cognitive strategies analogous to computer simulation when intuitively reasoning and making predictions about the physical world (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Hamrick, Smith, Griffiths, & Vul, 2015; Smith, Dechter, Tenenbaum, & Vul, 2013; Smith & Vul, 2017; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). In order to account for the more imprecise and qualitative nature of human physical reasoning, they propose that multiple simulations are run from a range of different initial configurations. For instance, consider a person asked to predict whether a tower of wooden blocks will fall over. A *probabilistic* mental simulation begins by assuming that each observer has an imperfect perception of the positions of the blocks (i.e., their precise locations in physical space) owing to perceptual limitations and occlusion. Based on this uncertain percept, the simulator samples a number of slightly different towers, each altered according to random (perceptual) noise. According to the theory, a reasoner might start with, for instance, ten initial towers and run a (possibly noisy) physics simulation forward until some termination point with the resulting outcomes driving their stability judgment. For example, if 8 of the 10 simulated towers fall over then a reasoner might estimate a 0.8 probability that the structure is unstable (Battaglia et al., 2013). We refer to this approach as "probabilistic mental simulation theory" (PMST).

PMST has been found to approximate human judgments in a diverse set of tasks, including judging how a 3-D tower of blocks will collapse (Battaglia et al., 2013), predicting the destination of a virtual ball on a 2-D bumper table (Smith et al., 2013), and predicting the proportion of a poured liquid that will end up on either side of a divider (Bates, Yildirim, Tenenbaum, & Battaglia, 2015), among others. However, this theory has been contested (cf., Davis & Marcus, 2015, 2016). Criticisms of this theory include the incompatibility of an accurate physical simulation engine with decades of psychological work documenting human errors in simplified physical reasoning tasks (Hegarty, 2004; Kubricht, Holyoak, & Lu, 2017; McCloskey, Caramazza, & Green, 1980; Proffitt, Kaiser, & Whelan, 1990; Siegler, 1976). In addition, in many situations, simulation would be computationally inefficient or impossible. For instance, if a closed can half full of sand is shaken, simulation would require calculating all the collisions of all the grains of sand (e.g., Kubricht et al., 2016) but if the goal is just to predict whether the sand remains in the can, that can be done through the application of a simple rule (Smith et al., 2013)

The goal of the present paper is to provide a strong empirical test of PMST. We begin by describing three core tenets of PMST that transcend specific applications of the theory and make important testable claims about human physical reasoning. We then describe three novel experiments that test these principles by setting up pre-registered (here) edge-cases where we might expect the predictions made by PMST to fail.

Three key principles of probabilistic mental simulation theory (PMST)

An agent using a probabilistic simulation of the physical world to solve physical reasoning problems should adhere to the following three principles. While we accept that PMST may include limitations and shortcuts (Ullman et al., 2017), the principles outlined here are necessary for simulation to be a viable strategy.

Object Persistence A reasoner using PMST is required to maintain interacting objects within all simulations/samples. Objects occupy particular locations within space and time and a mental simulation must encode these relative spatiotemporal positions and update them according to the rules of physics. This is a core aspect of the theory, because drop-

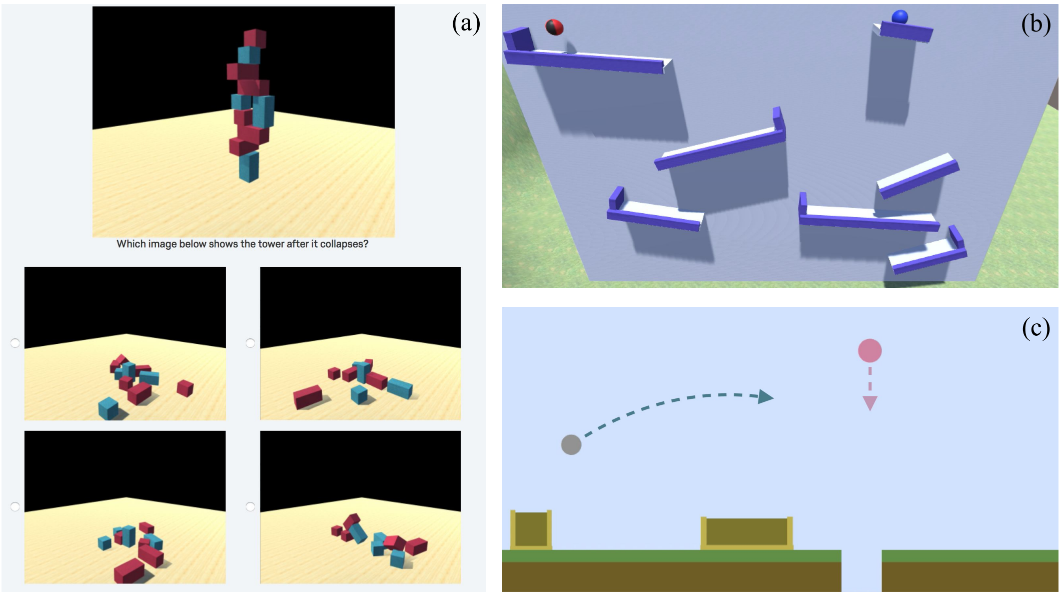

Figure 1: (a) A "Minus One" block tower question, as it appeared to participants. The answer in the upper left is correct; the other three answers are each missing one block. (b) A "marble run"-type temporal consistency question. (c) An example probabilistic coherence scene. Dotted arrows indicate approximate motion over the 2/3 second clip.

ping an object from a simulation or deleting it (as is possible in a video game engine) would radically alter possible outcomes and subsequent predictions. For example, imagine a person thinking about a table. If their mental simulation accidentally deleted its representation one of the table legs (even temporarily), the result would be a major disruption to the simulation; e.g., the otherwise static table might begin to fall. Keeping track of the location of objects in space and time and accounting properly for their movements is fundamental to what it means to "simulate" a physical scene. Any plausible physics engine must keep track of all the interacting objects involved in order to maintain coherent predictions about the future. Physics engines do make mistakes and approximations, but deleting or radically altering objects is not the kind of mistake they make.

There are, of course, some cases where objects may be ignored. In video game physics engines, for example, objects at rest are often put to "sleep" to save on computation, and it has been suggested that mental simulation might make use of this trick as well (Ullman et al., 2017). When a physics engine puts an object to sleep, however, this simply means that the engine assumes that the object is stationary, and it does not mean that the physics engine forgets that the object exists.

This leads to the key prediction tested in our first experiment: in interacting multi-object scenes, every object from the initial percept will be represented in each simulation's final state, because every object is necessarily represented and tracked throughout each simulation.

Temporal Consistency Building upon the first principle, an iterative simulation must advance all interacting objects simultaneously. This step-by-step, synchronous nature of the simulation ensures that the order of events is preserved. Preserving the order of events is important for generating accurate simulations and using them to make decisions. When two processes might interact, it is necessary for their simulations to be properly synchronized in order to predict whether and how they interact. Consider a case where a bottle is rolling towards the front door of a house, which is slowly closing. To predict if the bottle ends up inside or outside the house, the simulator has to represent whether the door will swing shut before the bottle gets there. Time cannot for instance run faster for the bottle than for the door if reasoning is to be coherent. A synchronous approach does this, and ensures that there is no way for one event to get ahead of, or fall behind another, because they share a common timeline.

We refer to this property of mental simulation as *temporal consistency*. A person using PMST to reason about a physical scene should preserve the temporal order of events.

Probabilistic Coherence According to PMST, after running multiple (noisy) simulations, the final scene configurations from each simulation are used to make predictions and inferences about the physical world. A variety of ways of aggregating across these simulations have been proposed. For example, in Battaglia et al. (2013), the output of the model was the average proportion of towers that fell across the set of the simulations. In this example, PMST uses the Monte Carlo principle to estimate probabilities. An event that is almost certain to occur will occur in all the simulations while a more uncertain event (or one more sensitive to perceptual noise) will occur less frequently. Although approximate, computing probabilities from samples or simulations still conforms to the axioms of probability theory. Indeed, this is a key virtue of the approach, and helps to relate the theory to existing Bayesian theories of human inference.

One classic signature of coherent probabilistic reasoning is that the probability of a conjunction of multiple events must always be less than the probability of any component (i.e., $P(A \wedge B) \leq P(A)$). However, people in many cases will estimate conjunctions to be more likely than one of their components (known as the "conjunction fallacy", Tversky & Kahneman, 1983). While they have been found in a number of domains including social reasoning, conjunction fallacy errors have not (to our knowledge) been observed in reasoning about physical outcomes, and all of the methods PMST proposes for estimating probability from the results of simulation predict that conjunction fallacies should not regularly occur. If likelihood is calculated by tallying the relative outcomes of different random simulations, the conjunction rule will not be systematically violated, because it is impossible for a sample to have more outcomes that include a conjunction than outcomes that include one of the constituent elements.

Study 1: Object Persistence

The first experiment tested the principle of Object Persistence. In particular, we tested if people are able to keep track of the number, size, and color/identity of a relatively small number of objects when predicting the future state of a simple scene. If people fail consistently at this task, it calls into question a key assumption of PMST; that simulations preserve objects over time. The assumption that we make in designing this test is that if an object is represented and tracked in each simulation, then it should be available for other judgements such as being identified/recognized. If people are limited in this regard, it calls into question if people use PMST or, alternatively, refines this theory by pointing out the cognitive inaccessibility of object-level details from the mental representation of the scene.

The experiment builds upon the block tower designs first used by Battaglia et al. (2013). Rather than asking participants to make predictions about the collapse of a standing tower, we showed participants one standing tower (the target) and then a set of four collapsed towers and asked them to judge which collapsed option was the result of the target tower falling according to gravity (with one of the four options being the ground truth of running the standard tower through a physics engine). Given the target tower, a simulation based reasoner could simply simulate the standing tower forward to generate one or a set of collapsed tower states. The actual result of the tower collapsing should be similar to several results generated by the simulation.

Method

Participants We ran groups of 9 at a time until the number of participants who meet the criteria reached or exceeded the planned number of participants, which was 100[1]. We recruited 201 participants (71 female, mean age = 33.9, SD = 9.8) on Amazon Mechanical Turk (AMT). Participants could earn a bonus of $3 depending on the accuracy of their predictions. Of these, 101 participants were eligible for our analysis. We analyzed the first 100 (39 female, mean age = 34.0, SD = 10.0). This collection plan and all criteria were outlined in our preregistration (here).

Stimuli The stimuli were still images of standing but unstable block towers (targets). Each target tower consisted of 10 blocks, similar to what has been used in previous research (e.g. Battaglia et al., 2013; Hamrick et al., 2016). The blocks came in three colors (red, blue, and green) and in three dimensions (the "cube" in 1x1x1, the "brick" in 1x1x2, and the "plank" in 1x0.5x2; units are relative).

For each target tower, there were four still images of possible resting states, i.e., what the tower might look like once it had collapsed under gravity. One of the resting states was always the real result of the target tower collapsing in the physics engine we used to create the stimuli.[2] The other three were incorrect and impossible in one of the following ways. In "Change Type" questions, one of the blocks was replaced with a block of different dimensions. In "Change Color" questions, one block was switched to a different color. In "Swap Color" questions, two-color towers swapped the colors of all blocks; e.g. all red blocks would become blue and all blue would become red. In "Plus One" questions, an additional block was included. In "Minus One" questions, one block was missing (e.g. Figure 1a). In "Minus Two" questions, two blocks were missing. In "Minus Three" questions, three blocks were missing.

The impossible endstates were created by changing the original tower (e.g. deleting, adding, or changing the properites of one or more blocks), adding some noise (so that all the incorrect answers were not identical), and then allowing the simulation to run to rest. Materials were created until there were three impossible endings that had no blocks that fell outside the viewing area nor were entirely obscured by other blocks.

Procedure Participants read a detailed description of the task. This included several example videos generated from the PhysX materials, and example images like those that appeared in the main body of the task. Participants were asked to watch each video a few times so that they would know how the blocks act when they fall.

The main body of the study consisted of 14 4AFC trials randomly intermixed with 10 easy trials. The easy trials were designed so that the correct answer would be obvious to a participant who was paying attention. Trial order was randomized. When choosing between the four fallen towers the original tower of blocks was still visible on the screen (see

[1] In an earlier preregistration (here), we allowed for a small number of exclusions. However, when we began collecting data for this study we realized that the exclusion rate was much higher than expected. As a result, we stopped data collection and developed a new protocol with a fixed n per experiment *after* exclusions. See Kennedy, Clifford, Burleigh, Waggoner, and Jewell (2018) for discussion of why the exclusion rate may have been unusually high during the summer of 2018, when the majority of these data were collected. In addition, due to space limitations we can report only the key planned analyses in this conference paper.

[2] The PhysX physics engine, through the Unity interface (*Unity*, n.d.).

Figure 1a).

Results

In accordance with our preregistered analysis plan, we pooled the number of correct answers participants gave on the 14 critical items, and used both a two-tailed one-sample *t*-test and the one-sample *"Bayesian Estimation Supersedes the* t*-Test"* or BEST (Kruschke, 2013) to estimate credible intervals for overall performance. The average number of correct answers was 6.33 (SD = 2.37). We calculated a 99% confidence interval of [5.75, 7.00], and the one-sample BEST gave a 99% credible interval of [5.72, 6.98]. Performance at this simple physical reasoning task is thus exceedingly poor; this contrasts sharply with the high performance at predicting whether towers are unstable found by Battaglia et al. (2013)

These errors varied by trial type. The mean number correct (out of 2) were 0.38 for Change Type items, 0.79 for Change Color items, 0.59 for Swap Color items, 0.65 for Plus One items, 1.18 for Minus One items, 1.36 for Minus Two items, and 1.38 for Minus Three items. We calculated confidence intervals corrected for multiple comparisons (Bonferroni with .05/7 = 0.00714) for all items. Intervals for Swap Color and for Plus One were consistent with a null of 0.50. For these item types, participants perform as poorly as if they were given no information at all. The 99.29% interval for the items with the highest accuracy, Minus Three, was [1.19, 1.59], the upper limit being just less than 80% accuracy. Notably, 39 of the 100 participants gave the correct answer to fewer than half of the items. Only 3 participants made no errors at all.

We included a free-response question after all trials, asking participants: "Roughly speaking, how did you try to solve the problems? Please tell us a little about your approach below." Three coders who had not been involved in the design of the study or the collection of data coded the free responses into the following categories: 0) No response, Nonsensical response, or "Other" strategy, 1) Simulation, Visualization, or Imagination, 2) Heuristics or Rules, 3) Both Simulation & Heuristics. To conduct subgroup analyses, we used a best 2 out of 3 approach to resolve disagreements among the coders, and had the three coders manually resolve disagreement for the small number of self-reports where all three coders coded the response differently. The ratings had a Cronbach's alpha of 0.85, indicating acceptable agreement (Kline, 2013).

When participants were asked to describe the way they completed the relevant tasks, 19 gave answers that suggested a simulation or visualization approach, 50 said they used specific rules or heuristics, 20 said that they used both simulation and heuristics, and the remaining 11 gave an uninterpretable answer. Results did not differ between participants who reported using different strategies.

Discussion

Reasoning about sets of 10 simple objects should be well within the abilities of a person using PMST (Battaglia et al., 2013; Hamrick et al., 2016). Despite this, performance was remarkably poor.

This behavioral result seems very unlikely if participants were tracking every block, which in causally-bound systems is a requirement of PMST. While it is possible that simulators might not always keep track of things like color, tracking shape is necessary to predict object interactions, and tracking every object is fundamentally necessary for the task. Because of this requirement, every object will end up in the end states of every simulation. It would seem trivial then to detect a mismatch between the end state of a mental simulation and a provided image of such a final scene. Alternatively, if one retained the spirit of the PMST approach, this result significantly constrains the availability of particular information within a mental simulation. Introducing this new constraint seems hard to reconcile with the ability of people to judge if the tower will fall via simulation because it would imply someone could answer the falling question ("will this block tower fall over?") but not a question about an individual block within a tower (e.g., "will the long red block remain standing when the tower falls over?").

Study 2: Temporal Consistency

PMST conducts simulations in an iterative fashion. At every time-step, the system applies elementary physical rules to each object in the simulation. This is done recursively; once every object has been updated at time *t*, the system moves on to time *t*+1, updates all objects again, and so on (Battaglia et al., 2013). This ensures that events will generally occur in the correct order, as long as the approximate trajectory is clear. In this study, we assessed if people have difficulty predicting the order in which events occur, for physical events with reliable trajectories.

The materials for this study consisted of video clips of events in a simple 3-D world. Participants viewed the first two seconds of several short clips of physical scenes in which two independent physical processes unfolded. For example, the physical processes might be two balls, each rolling down its own series of ramps (see Figure 1b), or they might be two lines of dominoes falling over. Each physical process followed a predictable trajectory, and we informed participants of this fact.

In each scene we identified one object in each process (usually "the red ball" and "the blue ball"), and participants were asked to predict which of the two objects would hit the ground first. Participants did not see the outcomes of the video clips, so they had to engage in prospective reasoning in order to make this judgment. The key dependent variable was the proportion of scenes for which participants thought the wrong event would occur first.

Method

Participants As above, our stopping rule was designed to collect a fixed number of participants *after* exclusions. We collected 78 participants (29 female, mean age = 35.1, SD = 9.8) in groups of 9 at a time on Amazon Mechanical Turk. Participants could earn a bonus of $3 depending on the accuracy of their predictions. Of these participants, 63 met our exclusion criteria, and we analyzed only the first 60 participants (22 female, mean age = 36.4, SD = 10.2), as stated in our preregistration.

Stimuli The main stimuli were video clips (example clip) showing the first two seconds of a scene (full version of same scene). Each scene included two key objects, one red and one blue, each involved in its own causal chain, which would eventually lead to each object colliding with the ground.

Each scene was designed to make the outcome that would occur second seem, at the end of the 2-second clip, more likely to occur first. The object that would actually strike the ground second was moving faster, had gone further, had fewer "obstacles" in its way, etc., or some combination of these factors. We iterated the design of the scenes based on these heuristics until we believed that pausing at the two-second mark would lead to incorrect judgment of the conclusion. PMST predicts that no such items should exist, as long as the trajectories are clear.

In the full scenes, the first object always struck the ground at least 2/3 of a second before the second one did, sometimes much earlier. The full scenes took about 10s to complete.

Procedure Participants read a detailed description of the task which included several example videos of the physics engine we used, and example clips similar to those that appeared in the main body of the survey. Participants were assured that the simulations were designed to be as much like real physics as was possible, that both critical objects would always eventually reach the ground, that there were no hidden objects or forces that would interfere, and that everything relevant to the scene was readily visible in the video clips.

In the main body of the study, participants viewed several video clips of the first two seconds of a physical scene where two independent chains of events unfold. In each case there were two items of interest, one red and the other blue, and participants judged which of the two would reach the ground (indicated by a grass texture) first.

The study presented four questions each of three types ("Marble Run", "Parthenon", and "Domino"), for a total of twelve critical questions. There were also four filler scenes, which were designed to be trivially easy.

Results

We used both a two-tailed one-sample t-test and one-sample BEST to determine if, on average, accuracy was different from chance. Participants answered a mean of 4.77 questions correctly (SD = 2.55), which was less than chance (6), according to both a t-test, $t(59) = -3.75$, $p < .001$, 95% confidence interval [4.11, 5.42] and a one-sample BEST, 95% Credible Interval: [4.11, 5.45].

In answering the 12 critical questions, 56.7% of the participants gave the incorrect answer to more than half of the questions. Every participant made at least two errors. The highest level of performance was ten of twelve correct, achieved by only two participants. Further, 3.3% of the participants gave the wrong answer on all twelve trials.

The same three coders coded free response reports of strategy according to the system described above. The ratings had a Cronbach's alpha of 0.74, indicating acceptable agreement (Kline, 2013). When participants were asked to describe the way they completed the relevant tasks, 8 gave answers that suggested a simulation or visualization approach, 35 said they used specific rules or heuristics, 8 said that they used both simulation and heuristics, and the remaining 9 gave no answer or an uninterpretable answer. Results did not differ between participants who reported using different strategies.

Discussion

In this study, participants saw two processes with predictable trajectories, and were asked to estimate which process would complete first. Overall, participants reversed the order of the events in their predictions, predicting that the event that truly occurred second would occur first, and did so more often than chance. Admittedly, the scenarios used in this study were deliberately designed to be adversarial. If we were to imagine the (hypothetical) space of all possible scenes, it is likely that few cases would prompt the reversals in judgment we observed. However, PMST suggests that *no items* showing such reversals should exist, barring major uncertainties in trajectory, etc. That there exist any items where this kind of reversal is consistently found is evidence that PMST is not the approach being used to make these judgments.

Study 3: Probabilistic Coherence

When making predictions about a physical scene, a key claim of PMST is that judgments reflect probabilistic inference, estimated via repeated stochastic runs of the simulation (Battaglia et al., 2013). As such, people's physical judgments should approximately obey the laws of probability theory.

Conjunction fallacy errors are cases where people rate a joint probability (A & B both occur) as more likely than the marginal probability of one component (e.g. A occurring at all). This is logically contradictory because there is no way for the joint probability to be larger than either of its components. At most, it will be equal to the smaller component.

In the cognitive domain, this is often known as the "Linda Problem", because of a well-known example in which participants judged a hypothetical individual named Linda as more likely to be both a bank teller and a feminist than to be a bank teller in general (Tversky & Kahneman, 1983). To test this commitment of PMST, in this study we assessed if people fall prey to conjunction fallacy-style judgment errors for physical reasoning problems.

Methods

Participants We collected data from 90 participants (28 female, mean age = 33.6, SD = 9.8) on Amazon Mechanical Turk (AMT). Following the criteria outlined in our preregistration, we analyzed only the first 60 participants (18 female, mean age = 34.2, SD = 9.7) of 62 eligible.

Stimuli The main stimuli were video clips, 2/3 of a second long, in which two round objects (a pink "sphere" and a gray "cannonball") interacted in a 2-dimensional world (example video here). This world included gravity and some stationary objects. There was always "ground" on the bottom edge of the scene, with a green section representing grass on top, and usually one or more boxes resting on the grass. There was always a hole in the ground, wide enough for either object to potentially fall into.

Over the course of each clip, the gray cannonball would travel in a parabola across the screen, while the pink sphere would fall under the influence of gravity (see Figure 1c). The cannonball always traveled toward the sphere, in a way that suggested that the two might collide. Each video ended after approximately 700 ms, well before the cannonball could intersect the pink sphere's path, leaving ambiguity about the outcome of the scene.

Procedure Participants read a detailed description of the task. This included several example simulation videos from the physics engine we used (PhysX) and example clips like those that appeared in the main body of the task. The example videos included many forms of inter-object interactions, including collisions, and participants were allowed to watch these videos as many times as they wanted.

In the main body of the study, participants saw several simple physical scenes. For each scene, participants estimated the likelihood of a particular prompted outcome (e.g., "How likely is it that the pink sphere will end up on the grass?"), as a percentage ranging from 0% to 100% in 1% increments.

Eight of the scenes were considered "critical", and the answers to these provided our primary dependent measure. Unknown to participants, each critical scene appeared twice, for a total of 16 critical trials.

For each scene that appeared twice, in one appearance participants were asked the question, "How likely is it that the pink sphere will end up on the grass?" and in the other, "How likely is it that the cannonball will hit the pink sphere, and then the pink ball will end up on the grass?" Scenes did not repeat until after several filler scenes were presented.

Results

We averaged the difference scores (conjunction probability - sole probability) for each participant for each of the eight critical scenes. Positive values on these difference scores indicate that participants rated a conjunction as more likely than the constituent sole probability, which is a form of the conjunction fallacy. The average rating difference score was 7.29 (SD = 13.07), which was reliably greater than zero, according to both a t-test, $t(59) = 4.32$, $p < .001$, 95% confidence interval [3.92, 10.67], and a one-sample BEST (Kruschke, 2013), 95% Credible Interval: [4.06, 10.79]. This suggests that, on average, participants were inclined to commit the conjunction fallacy in a physics domain.

In rating conjunction and sole probabilities on critical trials, 72% percent of the participants show a bias toward the conjunction event. In addition, 62% percent of subjects committed the conjunction fallacy for more than half of the pairs.

The same three coders coded free response reports of strategy according to the system described above. The ratings had a Cronbach's alpha of 0.75, indicating acceptable agreement (Kline, 2013). When participants were asked to describe the way they completed the relevant tasks, 23 gave answers that suggested a simulation or visualization approach, 17 said they used specific rules or heuristics, 12 said that they used both simulation and heuristics, and the remaining 8 gave no answer or an uninterpretable answer. Somewhat surprisingly, we found that participants actually made *more* extreme conjunction fallacy errors when they reported using a simulation approach, $F(3, 56) = 3.90$, $p = 0.013$.

Discussion

Participants making judgments about outcomes in physical processes routinely predicted that conjunctions were more likely than one of their constituent events. PMST states that judgments about the outcomes of physical processes are made by aggregating over the result of multiple noisy runs of a simulation, and so conjunction fallacy errors contradicts this aspect of the theory.

General Discussion

Simulation has been argued to be an important and effective way in which people reason about the physical world. In this paper we ask about the limits on the use of simulation as a strategy. Across three studies, we found empirical contradictions to the natural predictions made by PMST.

First, when trying to identify the resting state for an unstable tower of 10 blocks, participants have great difficulty distinguishing between the true set of blocks and sets that differ because of changes of color, changes of dimensions, additions, or deletions. PMST suggests that this should not happen without significant additional assumptions about the content and accessibility of particular features of simulated representations.

Second, when judging the order of events in a scene with highly predictable trajectories, participants consistently make incorrect predictions about the order of events. Although the examples were designed to be adversarial, PMST does not admit the existence of such examples because judgments are made using an iterative simulation where every object is advanced synchronously in each unit of time.

Third, participants consistently commit the conjunction fallacy (Tversky & Kahneman, 1983) when reasoning about simple physical scenes, a result that contradicts the claims of PMST about how estimated judgments of physical scenes are made by aggregating across probabilistic samples.

The design of our experiments tried to mimic many of the empirical studies which have supported PMST in complexity and content. Thus we believe they represent an interesting test bed for the generalization of the theory.

As the field tries to grapple with these complex questions, we argue that any complete account of human physical reasoning must contend with both the cases where people appear to do well and the situations where they apparently are limited or deceived. As a result, experiments exposing the limits of simulation can be as informative as those that show the successes.

Acknowledgments The authors thank Ellie Robbins, Michael Lepori, Xuechen Sheryl Zhang, and Adi Kwiatek for help with this research and Gregory L. Murphy for helpful comments.

References

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2015). Humans predict dynamics using probabilistic simulation. In *Proceedings of the 37th annual conference of the cognitive science society.*

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013, November). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

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

Davis, E., & Marcus, G. (2016, April). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016, December). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

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 *Proceedings of the 37th annual conference of the cognitive science society.*

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

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P., & Jewell, R. (2018). The shape of and solutions to the mturk quality crisis. *Unpublished manuscript*.

Kline, P. (2013). *Handbook of psychological testing*. Routledge.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017, October). Intuitive Physics: Current Research and Controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759.

Kubricht, J. R., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., & Lu, H. (2016). Probabilistic simulation predicts human performance on viscous fluid-pouring problem. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 1805–1810).

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Nave beliefs about the motion of objects. *Science*, *210*(4474), 1138–1141.

Proffitt, D. R., Kaiser, M. K., & Whelan, S. M. (1990, July). Understanding wheel dynamics. *Cognitive Psychology*, *22*(3), 342–373.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive psychology*, *8*(4), 481–520.

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over time. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Smith, K. A., & Vul, E. (2017). Thinking inside the box: Motion prediction in contained spaces uses simulation. In *Proceedings of the 39th annual meeting of the cognitive science society.*

Tversky, A., & Kahneman, D. (1983, October). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, *90*(4), 23.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017, September). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Unity. (n.d.). Retrieved from `https://unity3d.com`