Broken Physics: A Conjunction Fallacy Effect in Intuitive 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

June 23, 2020

Abstract

One remarkable aspect of human cognition is our ability to reason about physical events. This paper provides novel evidence that intuitive physics is subject to a peculiar error, the classic conjunction fallacy, where people rate the probability of a conjunction of two events as more likely than one constituent (a logical impossibility). Participants viewed videos of physical scenarios and judged the probability that either a single event or a conjunction of two events would occur. In Experiment 1 (n = 60), participants consistently rated conjunction events as more likely than single events for the same scenes. Experiment 2 (n = 180) extends these results to rule out several alternative explanations. Experiment 3 (n = 100) generalizes the finding to different scenes. This demonstration of conjunction errors contradicts claims that such errors shouldn't appear in intuitive physics and presents a serious challenge to current theories of mental simulation in physical reasoning.

1 Introduction

Successful interaction with our environment often requires us to estimate the likelihood of particular physical events. For example, when deciding whether to walk through a construction site, we might gauge the chance of being injured by a falling piece of scaffolding. Accurately assessing the risk requires us to estimate the probability that certain physical events might occur in the future (e.g., a bolt might come loose). Frequently, we might also need to judge the probabilities of occurrence of conjunctions of several events (e.g., a support tube bends *and* a bolt comes loose).

There are good reasons to expect that estimates of probabilities in everyday physical situations should be well calibrated and internally consistent. In contemporary simulation theories of physical reasoning (Battaglia, Hamrick, & Tenenbaum, 2013; Ullman, Spelke, Battaglia, & Tenenbaum, 2017) the probability of physical events is argued to be estimated by sampling from noisy simulations consistent with the known state of affairs. A key consequence of this claim is that these estimates should satisfy the constraints of probability theory contingent on the samples themselves. Finally, a growing body of experiments suggest that humans, even pre-verbal infants, are sometimes capable of probabilistic reasoning about physical situations (Téglás et al., 2011; Xu & Denison, 2009; Xu & Garcia, 2008).

Further, there are reasons to expect physical reasoning to be different than other forms of reasoning. Every one of our ancestors had to navigate the same physical world; the parameters are exceptionally stable; correct physical reasoning is particularly valuable in evolutionary terms. For these and other reasons, philosophers and cognitive scientists have argued that intuitive physics will be unlike other forms of commonsense reasoning. Based in part on results with infants, Strevens (2013) specifically conjectured that the well-known flaws people demonstrate in probabilistic reasoning (cf., Kahneman, Slovic, & Tversky, 1982) appear only in dealing with subjective likelihoods ("epistemic probabilities") and not in dealing with physical probability. Recent work (Firestone & Scholl, 2016, 2017) has argued that intuitive physics is similar to low-level, automatic perceptual processes. The aim of this report is to document and report a novel error in reasoning that represents a challenge to this view.

1.1 The Conjunction Fallacy

If a reasoner estimates the probability of physical event A (e.g., a bolt comes loose on some scaffolding) as $P(A)$ and the probability of physical event B (e.g., a support tube bends) as $P(B)$, logically the probability of both events occurring must be equal to or lower than that of either component occurring (i.e., $P(A \wedge B) \leq P(A)$ or $P(B)$). However, decades of research has revealed that for many described scenarios, people tend to rate a conjunction as more likely than one or both of its constituents (Tversky & Kahneman, 1982, 1983), a reasoning error known as *the conjunction fallacy*.

The classic paper first reporting the conjunction fallacy (Tversky & Kahneman, 1982) included evaluations of a woman named Linda who fit the description of a progressive (e.g., Linda was described as concerned with social justice and in opposition to nuclear weapons). Based on this description, participants responded that Linda was more likely to be a feminist than to be a bank teller. Surprisingly, 85% of participants also rated "Linda is a bank teller and is active in the feminist movement" as more likely than "Linda is a bank teller", a logical impossibility. A common explanation of this error is that the conjunction statement mentions a "representative" trait or event (being a feminist) which is rated as highly probable on its own, whereas the single trait (being a bank teller) seems less representative of the evoked stereotype.

The conjunction fallacy has been explored in numerous subsequent papers. For example, Tversky and Kahneman (1983) tested several variations on the Linda problem. This included replications with both between- and within-subject designs and tests on populations with different levels of statistical skill (including undergraduates, medical students, and decision science Ph.D. students), with all variations confirming the original result. The conjunction fallacy has, of course, received a great deal of theoretical and empirical scrutiny since its introduction (Fiedler, 1988; Gigerenzer, 1991; Hertwig & Gigerenzer, 1999), much of the scrutiny focusing on concerns around pragmatics. However, rigorous empirical work has provided continued support for the finding and its status as a genuine reasoning fallacy (Bonini, Tentori, & Osherson, 2004; Sides, Osherson, Bonini, & Viale, 2002; Tentori & Crupi, 2012).

1.2 Reasoning about Conjunctions of Physical Events

The conjunction fallacy has been examined with a range of different materials, including judgments of the traits of individuals, estimations of the likelihood of natural disasters, predictions about federal legislation, and medical diagnoses (Tversky & Kahneman, 1983). Despite this, to our knowledge it has never been documented in the domain of physical reasoning.

The potential existence of a conjunction fallacy in the domain of physical reasoning poses serious problems for theoretical accounts of physical reasoning based on mental simulation (Battaglia et al., 2013; Ullman et al., 2017). In these models, predictions about the likelihood of various events depend on examining one or more outcomes of a mental simulation which maintains an approximate isomorphism to the physical dynamics of the actual world, similar to the way video games approximate real physical dynamics. To estimate if a tower of blocks will fall over, such theories argue that reasoners form an approximate mental representation of the configuration of each block in the tower, run forward a number of mental simulations each from a slightly different starting point (owing to sources of perceptual uncertainty about the precise configuration of the starting state of the simulation), and make final judgments by aggregating over the results of these simulations.

Probabilities estimated using this type of Monte Carlo simulation necessarily conform to the axioms of probability theory. For example, the frequency with which A and B both occur in different randomly initialized simulations must be less than or equal to the frequency of either event occurring alone across those simulations. There is no way for A & B to occur without A occurring as well, meaning judgments made using relative counts across a sample of simulations to estimate probabilities will always avoid the conjunction fallacy. This consistency with the laws of probability is a key virtue of the simulation approach, enabling sophisticated forms of inductive inference (Ullman et al., 2017). Thus, irrespective of accuracy, these theories predict that there should be no systematic violation of the axioms of probability in subjective judgements about common physical scenarios.

Often it is assumed that the classical theory of probability is the correct method of representation, but there are other theories of probability. A body of recent work has suggested that cognitive models of judgments may be better fit by the more general quantum probability framework (Pothos & Busemeyer, 2013; Pothos, Busemeyer, Shiffrin, & Yearsley, 2017). These data may be revealing to this debate as well.

2 Experiment 1

To evaluate the possibility of a conjunction fallacy in the domain of physical prediction, we employed a within-subjects design in which each participant viewed a number of clips showing simple physical scenes in a 2-D world. Participants viewed the first few seconds of each scene and rated the probability of a future event occurring if the scene were to continue (e.g., "What is the probability the ball will fall in the hole?"). Rating the probability of specific future events is a common task that has been used in many recent papers on intuitive physical reasoning (Battaglia et al., 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Hamrick, Smith, Griffiths, & Vul, 2015).

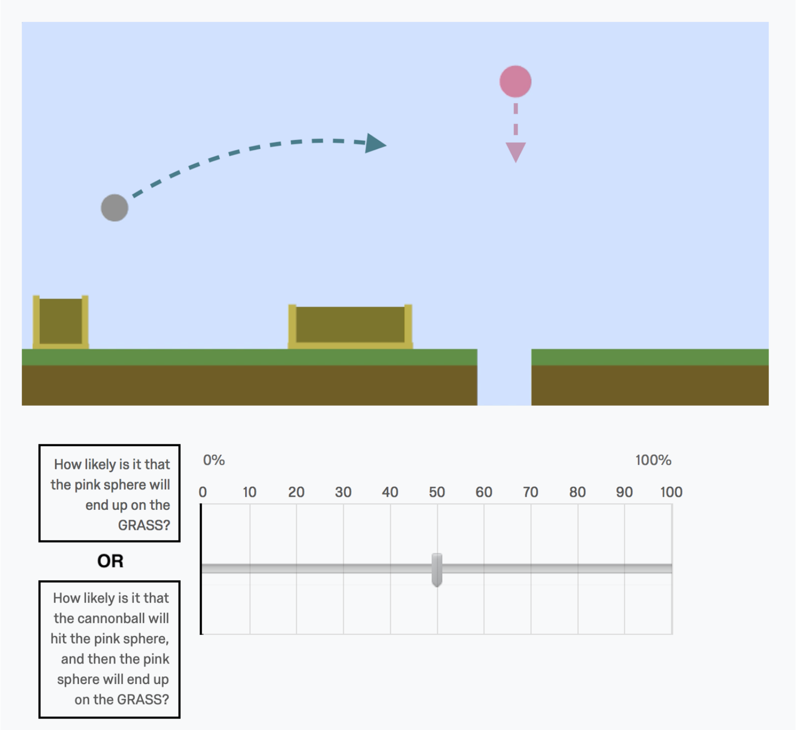
Each critical scene appeared twice, but participants were not informed of this fact. For these eight critical scenes, participants rated a conjunction event $P(A \wedge B)$ on one appearance and one constituent event $P(A)$ on the other. If participants rate the *conjunction* probability as more likely than the *constituent* probability, this is a form of the conjunction fallacy.[1]

2.1 Method

2.1.1 Participants

We recruited 90 participants (28 female, mean age = 33.6, SD = 9.8) on Amazon's Mechanical Turk. Of these participants, 74 were able to answer basic comprehension questions about the task, given three attempts. Following the exclusion criteria (outlined below), 62 participants were eligible for our analysis. We analyzed only the first 60 participants (18 female, mean age = 34.2, SD = 9.7), as stated in our preregistration (here). Pilot testing of the materials suggested a robust effect even at small sample sizes; based on the pilot data, we chose this sample size to provide high power ($> .95$) to detect the effect.

[1] A limited account of the results of Experiment 1 was previously published as a conference paper (Ludwin-Peery, Bramley, Davis, & Gureckis, 2019).



How likely is it that the pink sphere will end up on the GRASS?

OR

How likely is it that the cannonball will hit the pink sphere, and then the pink sphere will end up on the GRASS?

Figure 1: An annotated example of a scene as it appeared to participants. Dotted arrows indicate approximate motion over the 2/3 second movie clip. The grey circle was described as a "cannonball" and the pink circle as a "sphere". Each scene appeared twice, once with the constituent question, "How likely is it that the pink sphere will end up on the GRASS?" and once with the conjunction question, "How likely is it that the cannonball will hit the pink sphere, and then the pink sphere will end up on the GRASS?"

2.1.2 Materials and Procedure

Upon accepting the survey and consenting to participate, participants read a detailed description of the task. This included several example videos of the physics engine we used[2] and example clips like those that appeared in the main body of the survey. These examples included many forms of inter-object interactions, including collisions, and participants were allowed to watch these videos as many times as they wanted. Participants were informed of the nature of the clips and we explained how we wanted them to report their estimates of likelihood.

Participants then answered seven simple comprehension questions about the task, and were given three attempts to do so. If they were able to answer these questions correctly, they moved on to the rest of the study.

In the main body of the study, participants saw several scenes in which a pink "sphere" fell above a hole in a grassy field, and a gray "cannonball" traveled across the scene in such a way that it could potentially collide with the pink sphere (Figure 1). One object was called a cannonball and the other a sphere so that participants would be less likely to confuse them based on the written description. Pilot testing indicated that calling both objects by the same name caused confusion.

[2] The PhysX physics engine, through the Unity interface (*Unity*, 2005).

The scenes differed in several minor ways as well, including the exact speed and position of the objects, the size of the hole, and the presence or absence of one or more boxes on the grass. Each video stopped after approximately 700 ms, well before the cannonball could possibly intersect with the pink sphere's path, leaving ambiguity about the outcome of the scene. For each scene, participants were asked to estimate the likelihood of a particular outcome, and express that estimate as a percentage chance.

Eight of the scenes were critical scenes, the answers to which provided our primary measure. Each critical scene appeared twice, but participants were not informed of this fact. Half of the scenes that appeared twice were mirrored horizontally in their second appearance. 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?"[3] and in the other, "How likely is it that the cannonball will hit the pink sphere, and then the pink sphere will end up on the GRASS?" Scenes did not repeat until after several filler scenes were presented and completed. All video materials are available on the OSF.

For the filler scenes, participants were asked questions unrelated to the outcomes of interest used in the critical trials, such as the likelihood that the cannonball might end up in one of the boxes or in the hole. Pilot data indicated that separation by a few filler scenes was sufficient to prevent participants from explicitly recognizing the repetition.

Following the completion of the main body of the study, participants were asked to describe how they answered the questions in the main task using the following prompt: "Roughly speaking, how did you try to solve the problems? Please tell us a little about your approach below." We also asked several open-ended questions intended to determine whether or not participants had noticed that some of the scenes appeared twice with different questions. Finally, participants answered several demographic questions, gave free-response feedback, and were debriefed.

All data and materials are available on the OSF.

2.2 Results

The study included two questions with trivially obvious outcomes (e.g. the cannonball had already missed the sphere and couldn't possibly collide with it). When answering these questions, 12 (16%) reported that the near-certain outcome was less than 90% likely or that the near-impossible outcome was more than 10% likely, and were not included in our analysis, as defined in our pre-registration.

[3] Possible resting state locations were presented in all caps for clarity.

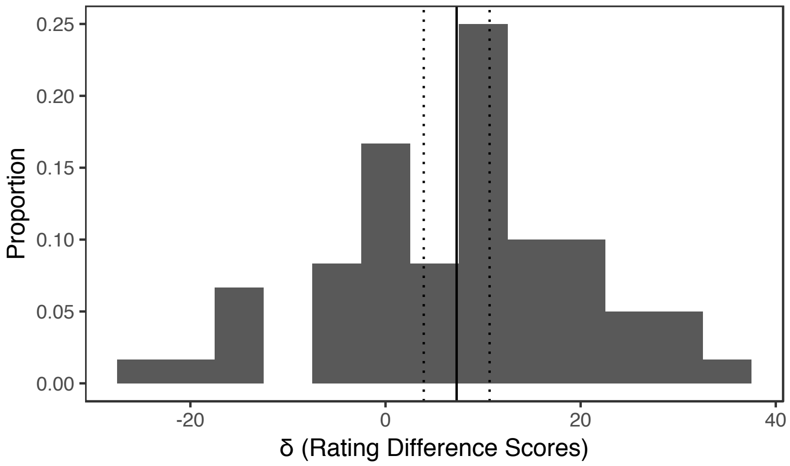


Figure 2: Distribution of the average of the 8 difference scores for each participant. Solid vertical line indicates the mean and dotted vertical lines indicate the upper and lower bounds of the 95% confidence interval.

2.2.1 Primary Analyses

For each of the eight critical problems, subjects rated the probability of the conjunction and sole statement. The mean judgment of the conjunction across all problems was 44.69% (SD = 25.88), and the mean judgment of the sole event across all problems was 37.40% (SD = 26.58).

The difference between these two ratings form our primary data. We averaged the rating difference scores (δ = conjunction rating − sole rating) for each participant for each of the eight problems (Figure 2). Positive values of δ indicate that participants rated conjunctions as more likely than their constituent sole events, which is a form of the conjunction fallacy. Zero or negative scores are not fallacious. The average δ value was +7.29% (SD = 13.07, SE = 1.69), which was reliably greater than zero, according to both a two-tailed one-sample t-test, $t(59) = 4.32$, $p < .001$, and a one-sample BEST, 95% Credible Interval: [4.06, 10.79]. The effect size was $d = 0.56$, 95% CI: [0.28, 0.83]. Therefore, participants appear to have erroneously rated conjunctions as more likely than their constituent sole events.

As shown in Figure 2, when averaging all critical trials together, 72% percent of the participants rated the conjunctions as more likely than their constituents on average. In addition, 62% percent of subjects rated the conjunction as more likely than the constituent on more than half of the critical pairs. If participants were respecting the laws of probability in their estimations, subjects would generally rate the conjunction as less or equally likely. We would certainly not expect to see a consistent reversal.

2.2.2 Secondary Analyses

In the post-experiment questionnaire, none of the participants reported noticing that some of the videos appeared twice. In a follow-up question revealing that some of the videos appeared twice, only seven participants claimed to notice. The questionnaire also asked participants to estimate how many videos were repeated. While the true number of repeats was eight, only three participants guessed close to this number. The majority of those who guessed said that only a small number (2 or 3) were repeated, though many participants declined to guess at all.

At the end of the study, we asked participants: "Roughly speaking, how did you try to solve the problems? Please tell us a little about your approach below."

Three coders who were had not been involved in the design or running of the study or the collection of data were asked to code the free responses into four categories. The ratings had a Cronbach's alpha of 0.75, indicating acceptable agreement (Kline, 2013). A one-way, between-subjects ANOVA found a significant effect, $F(3, 56) = 3.900$, $p = 0.013$, $\eta^2 = .17$, and a Bayesian test produced a Bayes factor of 4:1 in favor of a difference by reported approach. However, as this result failed to replicate in the higher-powered Experiment 2, we suggest that this result was a false positive and do not draw any conclusions from it.

3 Experiment 2

Experiment 1 provides evidence that people commit the conjunction fallacy when reasoning about simple physical scenarios. However, there is reason to be cautious in interpretation. Early criticisms of conjunction fallacy results centered on the idea that participants might be interpreting the task in line with conversational norms (Gigerenzer, 1991, 1996). People expect statements in conversation to be informative, truthful, relevant, and clear (Grice, 1991), but it can be argued that conjunction questions sometimes violate these expectations, leading participants to answer a slightly different question than the one intended. Various empirical attempts have been made to demonstrate that apparent conjunction fallacy errors are simply the result of participants reasonably misinterpreting the materials (Dulany & Hilton, 1991; Fiedler, 1988; Hertwig & Gigerenzer, 1999; Mellers, Hertwig, & Kahneman, 2001; Morier & Borgida, 1984), though even in the strictest tasks, the errors persist (Sides et al., 2002; Tentori & Crupi, 2012).

Two alternative interpretations seem particularly of concern in Experiment 1. Mellers et al. (2001) point out that, for example, "We invited friends and colleagues to the party" implies a union of friends and colleagues, rather than an intersection. If participants are reading the question quickly, they could potentially interpret P(A∧B) as something like P(A∨B) (i.e. "How likely is it that the cannonball will hit the pink sphere *or* the pink sphere will end up on the grass *or* both events will occur?”).

Similarly, participants might interpret the conjunction as the conditional (“*If* the cannonball hits the pink sphere, *then* how likely is it that the pink ball will end up on the GRASS?”). The conditional being larger than one constituent is not a logical impossibility, and certainly not a form of the conjunction fallacy.

To account for these classes of alternative interpretations, we ran a replication of Experiment 1 with additional conditions, allowing us to evaluate if alternative phrasings of the conjunction question would lead to the same conjunction errors observed in Experiment 1.

3.1 Method

3.1.1 Participants

We collected 269 participants (98 female, mean age = 35.3, SD = 10.2) on Amazon’s Mechanical Turk. Exclusion criteria were the same as in Experiment 1. We analyzed only the first 180 participants (60 female, mean age = 36.2, SD = 10.2), as stated in a new preregistration (here). This sample size was chosen so that each of the three conditions would have the same sample size as Study 1, 60 participants after exclusions.

3.1.2 Materials and Procedure

Experiment 2 was identical to Experiment 1, save for the phrasing used in the appearances of the conjunction question. We tested three different phrasings of the conjunction question, in three between-subjects conditions. The first was the **Original Phrasing** condition, in which participants saw the same phrasing that appeared in Experiment 1, namely:

> **Original:** “How likely is it that the cannonball will hit the pink sphere, and then the pink sphere will end up on the GRASS?”

We compared this to two alternative phrasings. The first was based on the original phrasing, but omitted the connecting adverb “then”. As a result, we call this the **No Adverb** condition. The comma was similarly omitted.

> **No Adverb:** “How likely is it that the cannonball will hit the pink sphere and the pink sphere will end up on the GRASS?”

The final condition was phrased to highlight the purely conjunctive nature of the question being asked. No causal language was used, and the components were presented in the reverse of what one would expect to be the natural order of events, so we call this the **Disordered** condition.

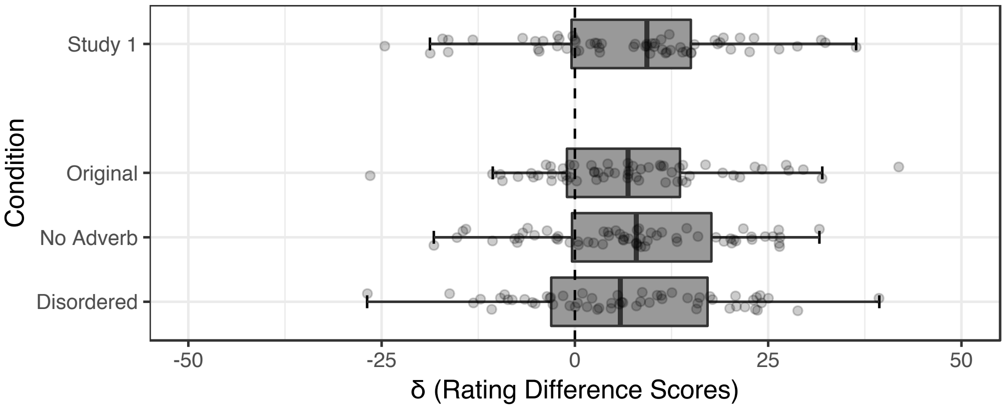


Figure 3: Distribution of the average of the 8 difference scores for the participants in each group. Included for comparison is the data from Study 1. Boxes correspond to the first and third quartiles.

> **Disordered:** "How likely is it that both will happen: the pink sphere will end up on the GRASS and the cannonball will hit the pink sphere?"

Participants were randomly assigned to one of the three conditions. Whenever they saw a conjunction question, they were given the appropriate phrasing by condition.

All data and materials are available on the OSF.

3.2 Results

The study included the same two trivially obvious questions as in Experiment 1. When answering these questions, 36 participants reported that the near certain outcome was less than 90% likely or that the near impossible outcome was more than 10% likely, and were not included in the analysis.

3.2.1 Primary Analyses

The mean judgment of the conjunction across all problems was 46.53% (SD = 28.20), and the mean judgment of the sole event across all problems was 39.35% (SD = 28.78).

As in Experiment 1, we averaged the rating difference scores (δ) for each participant for each of the eight problems (Figure 3).

Across all three conditions, the average rating difference score was +7.18% (SD = 12.34, SE = 0.92), which was reliably greater than zero, according to a two-tailed one-sample t-test, $t(179) = 7.81$, $p < .001$, and a one-sample BEST, 95% Credible Interval: [5.30, 9.00]. The effect size was $d = 0.58$, 95% CI: [0.42, 0.74]. Again, participants systematically rated conjunctions as more likely than a constituent sole event.

The effect of condition on rating difference scores was not significant, $F(2, 177) = 0.068$, $p = .93$, $\eta^2 = 0.001$, and Bayesian analysis found strong evidence of no difference, $\text{BF}_{10} = 0.067$. We dummy coded condition in a linear regression and in a Bayesian linear regression, the Original condition being used as the reference level, to compare the new phrasings to the Original phrasing. Neither slope was significantly different from zero, all p's $> .80$, all Bayesian 95% credible intervals for the slopes including zero, indicating no differences between each alternative condition and the Original condition.

None of the 95% confidence intervals on the slopes indicated differences that would reduce any condition to zero. Further one-sample t-tests showed consistent differences from zero, $p < .001$ for all three conditions, and all Bayesian 95% credible intervals not including zero.

3.2.2 Secondary Analyses

As before, three new coders coded free response reports of strategy according to the system described above. The ratings had a Cronbach's alpha of 0.78, indicating acceptable agreement (Kline, 2013).

A Chi-square test found no evidence of a relationship between condition and approach used, $\chi^2(6, \mathrm{N} = 180) = 9.56$, $p = .14$, and a Bayesian test of association produced a Bayes factor of 5:1 against a relationship between condition and approach, suggesting that the different phrasings did not influence choice of the approach used to solve the problems.

To determine if there was an overall impact of approach, a one-way between subjects ANOVA was conducted to compare the effect of reported approach on overall ratings on the critical items. There was no significant effect, $F(3, 176) = 0.582$, $p = .63$, $\eta^2 = 0.01$, and a Bayesian test produced a Bayes factor of 13:1 against a difference by reported approach.

4 Experiment 3

Experiment 2 establishes that the physical conjunction fallacy effect is highly consistent across alternative phrasings of the critical question, suggesting that the physical scenario itself, rather than the pragmatics of the question, produces this judgment pattern.

However, Experiments 1 and 2 investigate only one specific instance of the physical conjunction fallacy (two balls colliding mid-air). One concern is that some unexplored idiosyncrasy of our original design was responsible for this effect. By exploring whether we find this error in a range of more and less similar scenes, we round out the evidence for a physical conjunction fallacy. To that end, in Experiment 3, we designed and tested a range of different physical scenarios which have varying similarity to those in Experiment 1 and 2.

4.1 Method

4.1.1 Participants

We collected 198 participants (70 female, mean age = 36.89, SD = 10.66) on Amazon's Mechanical Turk. Exclusion criteria were the same as in Experiment 1. We analyzed only the first 100 participants (36 female, mean age = 38.34, SD = 10.17), as stated in a new preregistration (here), as this new experiment was added at the request of reviewers.

4.1.2 Materials and Procedure

We piloted a few different types of physical scenes before settling on three designs to develop further. These three new types of scenes were developed to investigate how widely the physical conjunction fallacy appears. As a result, the first new type of scene is somewhat similar to our original design, the second departs in certain critical ways, and the third is intentionally different along several axes. All materials are available on the OSF.

The first, "Tubes", is moderately similar in design to the original scenes (Figure 4a). Like the scenes used in the previous experiments, "Tubes" scenes involve two balls, both moving. As in our original scenes, one of the key questions is about whether or not the two balls will collide, and the other question concerns where one of the balls will end up. This type of scene always involves a red and a blue ball each exiting a tube, moving in the same direction, and flying towards a bucket, which they might land in but might equally overshoot. The sole question was always, "How likely is it that the BLUE ball will end up in the BUCKET?", and the conjunction question was always, "How likely is it that the RED ball and the BLUE ball will collide, and the BLUE ball will end up in the BUCKET?"

The second, "Basket Tube", represents more of a departure from our original design (Figure 4b). While these scenes also include a possible collision event, there is only ever one moving object, a single pink ball. This limits the amount of information participants need to keep track of to make an informed decision. The object in motion and its trajectory will always be the focus of attention, so different ways of asking the question shouldn't bring new sets of objects into scrutiny. Each scene showed the pink ball flying in an arc above a field with a grey tube sticking up out of it. Each scene paused at a point where the ball clearly would not fall into the tube if continuing on its parabolic trajectory. However, there was always a blue "backstop" ahead of the ball, which the ball could possibly hit. If the ball were to hit this backstop, it was possible that it might bounce off the backstop and into the tube. The sole question was always, "How likely is it that the ball will end up in the TUBE?", and the conjunction question was always, "How likely is it that the ball will hit the BLUE backstop and the ball will end up in the TUBE?"

The final type of new scenes, "Still Weight Tower", represents the greatest departure from our

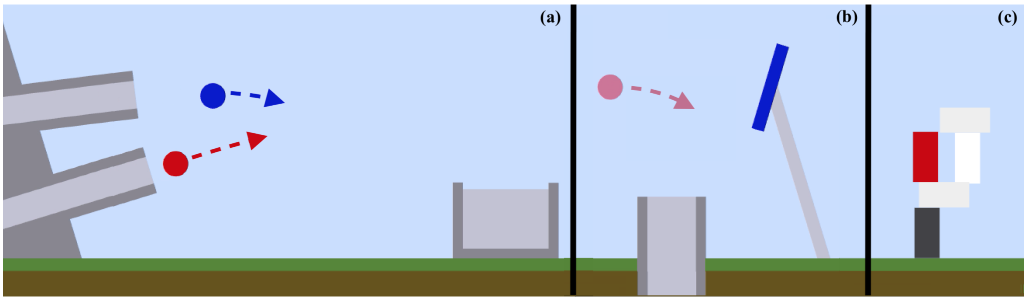


Figure 4: An annotated example of the three new scene types approximately as they appeared to participants. Dotted arrows indicate approximate direction of motion, if any. All scenes were in actuality the same dimensions, and are truncated here to conserve space. a) An example of a "Tubes" scene at the point where the video clip stopped. b) An example of a "Basket Tube" scene at the point where the video clip stopped. c) An example of a "Still Weight Tower" scene. In this type of scene, the tower was presented as a still image rather than as a short video clip.

original design, and therefore shows the greatest generalization of the effect (Figure 4c). These scenes were still images, rather than short videos; they involved no objects with initial motion; they involved the complex interaction of more than two objects, rather than simple collisions; and while in all other designs, both components of the conjunction are events, in this case one of the critical questions was about a parameter (the weight) of a particular object in the scene. Each scene portrayed a small standing block tower of 5 or 6 blocks. Some blocks were pale, some were dark, and one block was always bright red. To orient them to the scene, with every such image participants were told, "the DARK blocks are HEAVY, the PALE blocks are LIGHT, and the RED blocks might be either HEAVY or LIGHT." As part of the design, these red blocks of ambiguous weight were always placed at a location in the tower where the question of their weight might contribute seriously to the tower's overall stability. The exact sole and conjunction questions differed somewhat between the different towers, but the sole question always concerned whether or not the tower would stay standing, e.g. "How likely is it that the tower will STAY STANDING?" and the conjunction question always added a question about the weight of the red block, e.g. "How likely is it that the RED block is HEAVY and the tower will STAY STANDING?"

There were four scenes of each type, for a total of twelve new scenes. Each of these scenes appeared twice, once with the conjunction question and once with the sole question. Scenes were intermixed with a small number of filler questions to help prevent recognition of previous questions.

Materials were presented as in the previous experiments. In the main body of the study, all participants saw both questions for all of the scenes, and estimated the likelihood of the stated outcome, as in previous experiments.

Finally, participants answered several demographic questions, gave free-response feedback, and were debriefed.

4.2 Results

The study included the same two trivially obvious questions as in Experiments 1 and 2. When answering these questions, 10 participants reported that the near certain outcome was less than 90% likely or that the near impossible outcome was more than 10% likely, and were not included in the analysis.

4.2.1 Primary Analyses

As in Experiments 1 and 2, we averaged the rating difference scores (δ) for each participant for each of the four problems in each of the three new problem types. As a result, we ended up with a δ for each of the new scene types.

Positive values of δ indicate that participants rated conjunctions as more likely than their constituent sole events, which is a form of the conjunction fallacy.

Tubes The mean judgment of the conjunction across all four scenes of this type was 43.15% (SD = 19.04), and the mean judgment of the sole event across all four scenes of this type was 33.07% (SD = 19.00).

The average δ value for the "Tubes" problems was +10.08% (SD = 17.22, SE = 1.72), which was reliably greater than zero, according to both a two-tailed one-sample t-test, $t(99) = 5.85$, $p < .001$, and a one-sample BEST, 95% Credible Interval: [6.40, 13.00]. The effect size was $d = 0.59$, 95% CI: [0.37, 0.80].

Basket Tube The mean judgment of the conjunction across all four scenes of this type was 34.73% (SD = 16.10), and the mean judgment of the sole event across all four scenes of this type was 31.03% (SD = 13.80).

The average δ value for the "Basket Tube" problems was +3.70% (SD = 10.93, SE = 1.09), which was reliably greater than zero, according to both a two-tailed one-sample t-test, $t(99) = 3.39$, $p = .001$, and a one-sample BEST, 95% Credible Interval: [1.60, 5.90]. The effect size was $d = 0.34$, 95% CI: [0.14, 0.54].

Still Weight Tower The mean judgment of the conjunction across all four scenes of this type was 44.77% (SD = 13.34), and the mean judgment of the sole event across all four scenes of this type was 40.00% (SD = 14.85). This is strong evidence that participants are not misinterpreting the conjunction as the conditional. *Conditional* on the red block's weight being known, the overall likelihood would be well over 50%, which we don't observe here.

The average δ value for the "Still Weight Tower" problems was +4.77% (SD = 14.38, SE = 1.44), which was reliably greater than zero, according to both a two-tailed one-sample t-test, $t(99)$

$= 3.31$, $p = .001$, and a one-sample BEST, 95% Credible Interval: [1.80, 7.40]. The effect size was $d = 0.33$, 95% CI: [0.13, 0.53].

Overall, it appears that there is a reliable tendency to make conjunction fallacy errors in all three of the new scene types.

5 General Discussion

This paper reports three experiments showing that people rate conjunctive events as more likely than their constituents across a set of physical reasoning problems. Experiment 1 demonstrated the effect, Experiment 2 showed that the effect was robust to a range of alternative phrasings, and Experiment 3 expanded the finding to a wider range of physical scenes.

While the conjunction fallacy is well-established, the detection of a similar effect in physical reasoning is unexpected for several reasons. Many arguments have been made that intuitive physical reasoning is distinct from other types of cognitive activities and will be immune to these types of errors. In addition, common explanations evoked for the conjunction fallacy seem hard to apply in this case. For example, the concept of "representativeness", seems less relevant to the physical domain, as there isn't such a salient category or schema to activate.

One possibility is that the conjunction fallacy is a general phenomenon that occurs across many domains because it is an fundamental error in our judgment capacities. However, in followup work we have found that while magnitudes of conjunction fallacy errors are often correlated with one another (e.g., the size of an individual's error on a "Linda" problem correlates with conjunction fallacy errors on reasoning about dice), physics conjunction errors do not correlate with these other problems, suggesting a distinct and novel mechanism (Ludwin-Peery, 2020).

We argue that these results are additionally intriguing because they are unexpected given recent accounts of probabilistic mental simulation. However, such theories might be modified in light of these results. For example, rather than aggregating across multiple simulation runs in order to make a probabilistic inference, people might use some type of biased aggregation scheme which results in judgement errors (Zhu, Sanborn, & Chater, 2020). It remains to be seen if this biased aggregation approach can provide a simultaneous account of all the other documented phenomena in intuitive physical reasoning.

Importantly, past work has frequently acknowledged that there might be cases where heuristics were employed instead of mental simulation (Battaglia et al., 2013). If simulation is abandoned for certain problems, it suggests a control problem for the brain to determine when to adopt a simulation and when to use a heuristic. The present experiments provide an important waypoint about when simulation might be abandoned that may help inform such theories.

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

References

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

Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. *Mind & Language*, *19*(2), 199–210.

Dulany, D. E., & Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, *9*(1), 85–110.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, *50*(2), 123–129.

Firestone, C., & Scholl, B. (2016). Seeing stability: Intuitive physics automatically guides selective attention. *Journal of Vision*, *16*(12), 689–689.

Firestone, C., & Scholl, B. (2017). Seeing physics in the blink of an eye. *Journal of Vision*, *17*(10), 203–203.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, *2*(1), 83–115.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592–596.

Grice, H. P. (1991). *Studies in the way of words.* Harvard University Press.

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.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. *CogSci*.

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*(4), 31.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge university press.

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

Ludwin-Peery, E. (2020). Limited domain structure for conjunction errors. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*.

Ludwin-Peery, E., Bramley, N., Davis, E., & Gureckis, T. (2019). Limits on the use of simulation in physical reasoning. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269–275.

Morier, D. M., & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychology Bulletin*, *10*(2), 243–252.

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, *36*(3), 255–274.

Pothos, E. M., Busemeyer, J. R., Shiffrin, R. M., & Yearsley, J. M. (2017). The rational status of quantum cognition. *Journal of Experimental Psychology: General*, *146*(7), 968.

Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, *30*(2), 191–198.

Strevens, M. (2013). *Tychomancy.* Harvard University Press.

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.

Tentori, K., & Crupi, V. (2012). On the conjunction fallacy and the meaning of *and*, yet again: A reply to Hertwig, Benz, and Krauss (2008). *Cognition*, *122*(2), 123–134.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293.

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.

Unity. (2005). Retrieved 2018-06-20, from https://unity3d.com

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*(1), 97–104.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. *Psychological Review*.