The capacity limits of mental simulation

Halely Balaban[1,2,3] * Tomer Ullman[2]
halelyb@openu.ac.il, tullman@fas.harvard.edu
[1]Department of Education and Psychology, The Open University of Israel
[2]Department of Psychology, Harvard University
[3]Brain and Cognitive Sciences, Massachusetts Institute of Technology

Abstract

People have severe capacity limits when they track objects in direct perception. But how many objects can people track in their imagination? In eight pre-registered experiments (N=277 total), we examined the capacity limits of mentally simulating the movement of objects in the mind's eye. In a novel Imagined Objects Tracking task, we had participants continue the motion of animated objects in their mind up to a pre-defined point. When tracking one object in the imagination (Experiment 1a), participants gave estimations well in line with ground truth. But, when imagining two objects (Experiment 1b), behavior altered substantially: responses when tracking two objects in the imagination were fit best by the predictions of a Serial Model that simulates only one object at a time, as opposed to a Parallel Model that simulates objects in tandem. The serial bottleneck is not due to response/motor limitations (Experiment 2), and is reduced – but not eliminated – by adding extremely strong grouping cues (Experiment 3). Additional studies validate that the serial effect is not due to noise, exists in both realistic and hyper-simplified physics, is unaffected by motivation, and is found also for naturalistic occlusion (Experiments S1-S4). Altogether, we find that the capacity of moving entities in the imagination is likely restricted to a single object at a time.

[*] Corresponding author. E-mail: halelyb@openu.ac.il

Introduction

There's only so much we can hold in mind. A well-studied example is the limited ability to track objects in a visual scene. Numerous studies using the Multiple Object Tracking paradigm (MOT; 1) have tested how well people track objects as they move about, and found that tracking is limited to a handful of objects (e.g., 2; 3; 4; 5), with ongoing, important debates regarding the exact limitations and their origins (e.g., 6; 7; 8; 9). These limitations have been examined in great detail in direct perception, but what if the objects are not moving in front of one's eyes, but in the mind's eye? What are the limits of moving objects in the imagination?

People's tracking of objects extends beyond immediate perception, though the exact dynamics of tracking unseen objects or predicting future paths is still debated. In the MOT paradigm, several studies have suggested that people do not extrapolate trajectories to track occluded objects (e.g. 10; 11), at least under most conditions (see 12, for an exception), and instead use heuristics. On the other hand, a main current line of research suggests people use 'mental simulation' to engage in physical prediction or inference, proposing that people continue the trajectories of objects step-by-step in their imagination (13; 14; 15; 16). This approach has accounted for how people reason about the dynamics of objects in a variety of cases (e.g. 17; 18). While there are ongoing discussions about people's deviation from pure simulation (19; 20; 21), here we take as a starting point the idea that people can and do mentally simulate the movement of objects – and use this process to predict, keep track of, and reason about the motion of bodies – but also that this simulation is limited. Given this starting point, our goal was to test whether imagining the future trajectories of objects can be done for more than a single object at a time.

Compared to the large volume of research that examines the capacity limits of processing information available to direct perception, little is known about the limits on tracking imagined objects. While important recent research on mental imagery has started to demonstrate that adding more objects to an imagined static scene increases task difficulty, as reflected in people's subjective reports and precision (22; 23), it does not determine the capacity limits of simulating object dynamics in imagination. To examine this, we developed a novel Imagined Objects Tracking task. In this task, people watch animated scenes in which objects move up to some pause-point. People are asked to continue the motion of the objects in their imagination, and judge the timing of various outcomes. We focused on timing, as opposed to other dependent measures such as location accuracy (which has been extensively examined and validated in previous work on intuitive physics, but does not determine capacity limits), for two reasons: this avoids imposing serial response requirements, and leads to quantitatively and qualitatively distinct predictions in models of varying capacity.

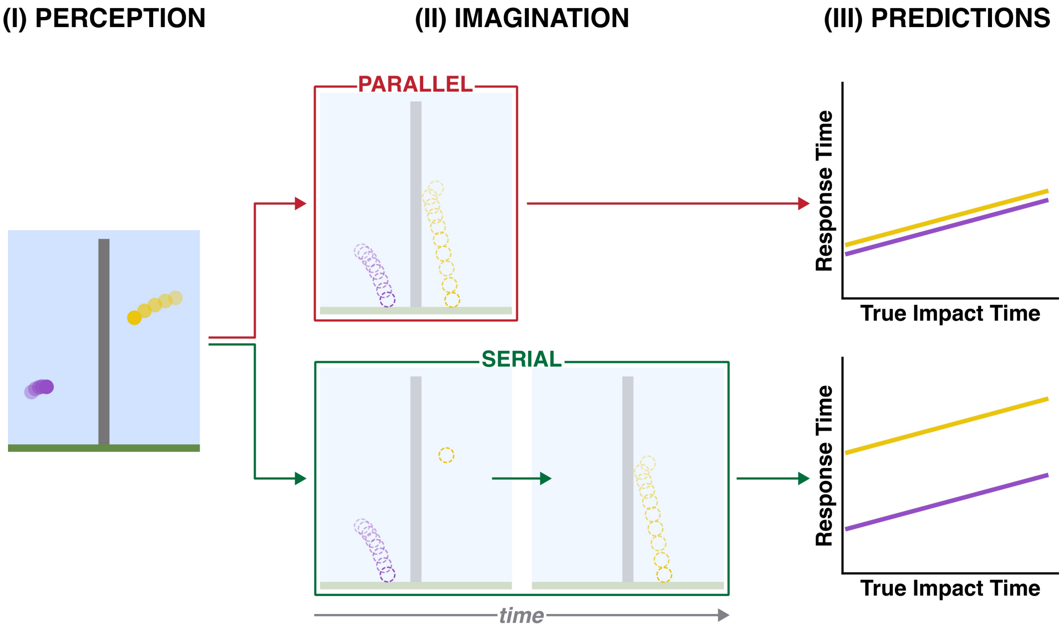


Figure 1: Theoretical overview. When watching a scene (I), people can track a handful of objects. But what is the capacity limit of moving objects in the imagination? In the Imagined Objects Tracking task, people watch animations of moving objects that pause mid-motion, and are asked to imagine how the motion continues and estimate the timing of outcomes – here, the moment when each ball hits the ground. Response times (reflecting the subjective impact time) are examined against the actual impact time, which can be manipulated. (II) A priori, moving objects in the mind's eye could happen in Parallel (top), with some number of objects moved forward simultaneously, or Serially (bottom), with only a single object advanced at a time. The Parallel and Serial Models make distinct predictions (III) regarding how people would assess the subjective impact time of objects in a dynamic scene. In the specific example shown in the figure, the Parallel model predicts the subjective impact time of both balls would be roughly the same, while the Serial model predicts a noticeable difference between the first ball moved forward in the imagination (here, the purple ball) and the second (here, the yellow ball).

We compared people's performance in Imagined Objects Tracking to two computational models that implement different hypotheses about the capacity limits of mental simulation (see Fig. 1). According to the Parallel Model, people can mentally advance multiple objects simultaneously. According to the Serial Model, people only advance a single object at a time, unfolding the trajectory of one before going back to unfold the trajectory of another. The Serial Model predicts that every additional object *differentially* increases the overall imagination-tracking time, delaying people's response for objects that are advanced later mentally. We note that several different sub-types of Serial Models are possible: people might simulate the motion of one object for a number of steps S, then switch to another object, then cycle back again to the first.

While a Serial Model that moves each object for a few steps at a time may appear a priori as an appealing solution to how people should mentally simulate objects, we find it completely deviates from the data in all of our studies. Furthermore, while an interleaved model might seem to be a middle ground between a fully serial model and a parallel one, its quantitative predictions do not reflect anything like an averaging of two 'extremes'. Because the interleaved model so clearly does not match our data, and because of its unintuitive predictions, the main text focuses on the serial model that first completely moves one object before turning to the next, but see the Supplementary Information, and the Discussion, for a complete analysis and consideration of interleaved serial models. We stress that both the Parallel and Serial Models 'keep around' the same number of objects. The capacity limit we studied is with regards to the mental simulation of the dynamics of the objects, and it is *not* the case that the Serial Model neglects the existence of an object when moving the other forward in time.

In eight pre-registered experiments, we studied the capacity limits of people's ability to mentally simulate the future paths of objects. As a benchmark, we first tested how precisely people track the timing of the imagined trajectory of a single object (Experiment 1a). Next and most important, we examined people's tracking of two objects in the imagination (Experiment 1b), and compared their behavior to the predictions of Parallel vs. Serial mental simulation models. We then further examined whether response requirements uniquely contribute to capacity limits (Experiment 2), and how scene regularities might help overcoming capacity limits in imagination through grouping (Experiment 3). In supplementary experiments, we validated that our results are not the effect of noise (Experiment S1), complex physics (Experiment S2), motivation (Experiment S3), or any unnatural disruption from freezing (Experiment S4). Our main finding from these studies is that people's capacity for moving objects in the imagination is extremely limited. Even in the minimal case of continuing the paths of two simple objects, people could only simulate the motion of one object at a time.

Results

Experiment 1a: Tracking a single object in the imagination

Participants in Experiment 1a saw animations of a single ball moving according to simulated physics, and pausing mid-motion. They were asked to continue the movement of the ball in their mind's eye and to press a key when the ball (in their imagination) hits the ground (Fig. 2, left). We compared participants' response times – indicating their subjective time estimation – with the actual time it would take the ball to hit the ground, based on the physical simulation. In different animations, the ball moved either like a cannonball or towards the wall, and the true impact time of the ball was manipulated by changing its height and velocity, producing values of 1, 1.2, 1.4, and 1.6 seconds from animation onset (see the Methods section for more details). The goal was to establish whether participants could imagine the future path of a single object in a temporally precise way.

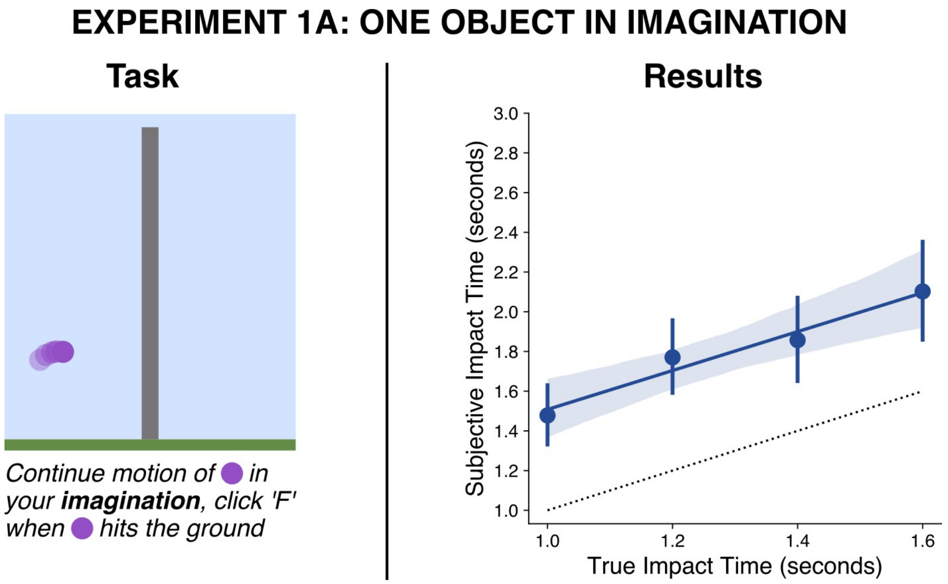


Figure 2: Task and Results of Experiment 1a: tracking a single object in the imagination. Circles indicate mean responses for different true impact time, error bars show SEM, solid line shows best linear fit, shaded area is 95% CI, and dotted line shows hypothetical perfect performance (where the subjective impact time equals the true impact time), as reference

.

As Figure 2 (right) shows, responses were linearly modulated by the true impact time (slightly lagging), F(1.73, 60.49) = 55.01, $p < 0.001$, partial $\eta^2 = 0.61$; linear trend: t(105) = 12.61, $p < 0.001$. Given that the delay is constant and is not modulated by the true impact time, we take the additive factor to reflect processes unrelated to the imagination component that is our focus, such as motor planning. The linear trend was not an artifact of averaging across participants, and can be seen in the individual data sets of almost all participants (see the Supplementary Information). The results suggest that people can indeed track the dynamics of a single object in their imagination when they observe scenes like those in our studies. These results serve as the basis for our critical question, which we tackled in the remaining experiments: what happens to people's ability to track objects in the imagination as more objects are introduced.

Experiment 1b: Tracking two objects in the imagination

Experiment 1b was identical to Experiment 1a, except that each scene included two objects (by combining two motion paths from two different scenes in Experiment 1a into a single scene), and the task was to press a different key when each object hits the ground (Fig. 3, top left). Scenes were created by combining two balls (one moving like a cannonball and one moving towards the wall) from the animations of Experiment 1a, with the true impact time determined independently for each ball, producing a true difference of either 0, 0.2, 0.4, or 0.6 between them. Again, we compared participants' subjective estimation of impact time with the ground-truth impact time, and also broke down their responses by order, meaning the first key press vs. the second one (see also the Supplementary Informationfor an analysis that focuses on the variation in responses instead of the means). Participants overall performed the task well (Fig. 3, top right), with a linear modulation of subjective impact time by true impact time, F(1.51, 52.86) = 34.99, $p < 0.001$, partial $\eta^2 = 0.5$; linear trend: t(105) = 10.04, $p < 0.001$. However, the second response happened much later than the first, F(1, 35) = 103.71, $p < 0.001$, partial $\eta^2 = 0.75$. The average delay was 640 ms (CI for the intercept of the first response: [618, 845] ms, second response: [1,032, 1,405] ms), and the interaction between response order and the true impact time was not significant, F(2.5, 87.43) = 2.06, p = 0.12. We note that the additive delay in response was smaller than in Experiment 1a, which might reflect a corrective attempt people engage in (i.e., speeding up the simulations to 'catch up' with reality), either explicitly or implicitly. Furthermore, the slope of responses is shallower than in Experiment 1a, suggesting participants are overall less tuned to subtle differences in ground truth physics, likely because of the harder task demands. Because these issues are independent from our main focus of a potential capacity limit in simulation, we set them aside as a target for future research.

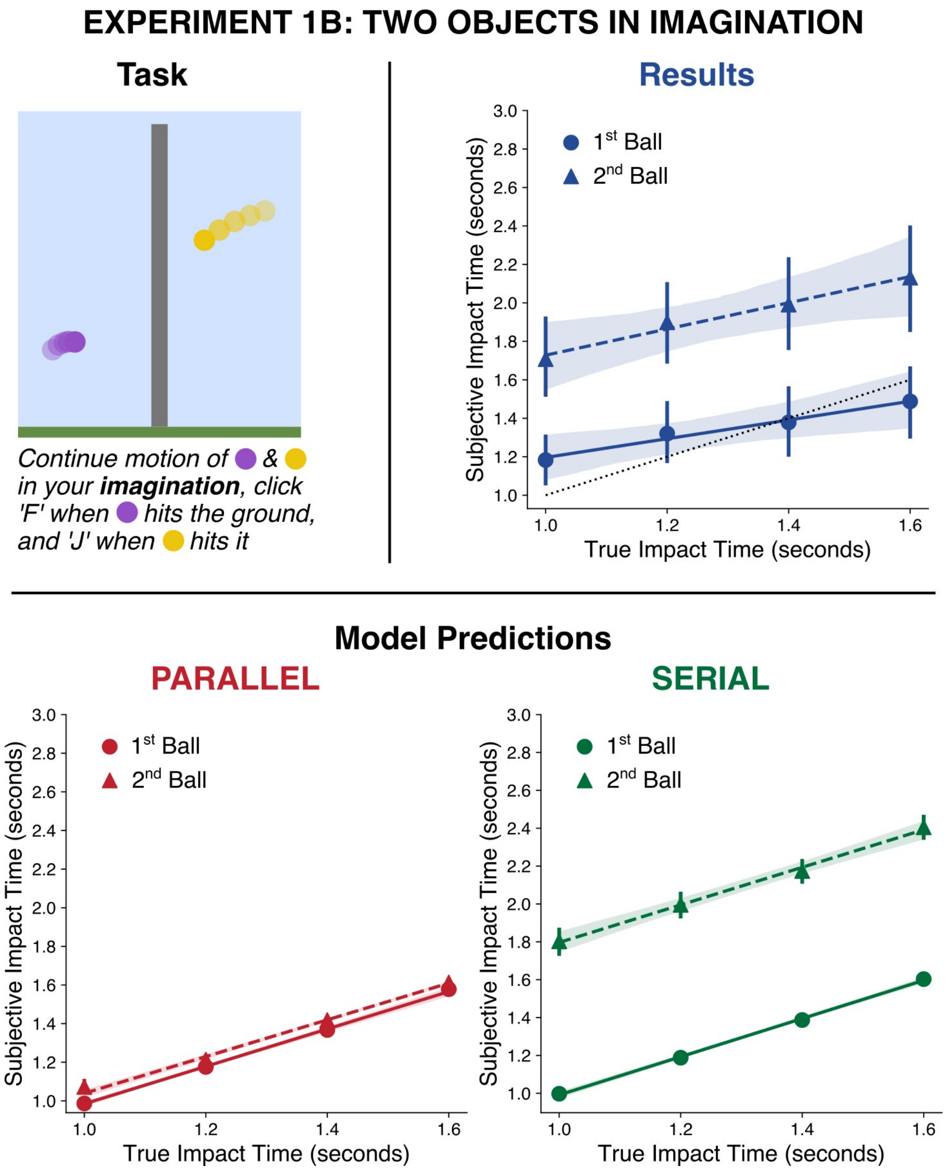


Figure 3: Task, results, and model predictions of Experiment 1b: tracking two independent objects in the imagination. Circles and triangles indicate mean responses for different true impact time and response order ('1st ball' refers to the ball participants responded to first, and '2nd ball' to the ball they responded to second), error bars show SEM, colored lines show best linear fit, shaded area is 95% CI, dotted line shows hypothetical perfect performance (where the subjective impact time equals the true impact time), as reference. Both models made similar predictions for the first response. For the second response, the Parallel Model predicted a minimal delay, due to perceptual noise. The Serial Model predicted a large delay for the second response, due to ending the simulation of the first object before turning to the second.

We created two simulation models of imagination tracking, using the physics engine that generated the stimuli. The Parallel Model produces two responses that are very close to each other, because both balls are advanced simultaneously (Fig. 3, bottom left). Only a small difference is expected, due to random noise in the simulation, which makes one random ball slightly faster on each run. Conversely, in the Serial Model the first ball has to run all the way through before the second ball can be simulated. Therefore, this model produces a large delay between the first and second responses (corresponding to the first and second ball to be simulated, respectively), in the order of several hundred ms (Fig. 3, bottom right). This is exactly what was found in our participants' data, as reflected by model fits: the Serial Model explained 96% of the variance in average responses, MSE = 0.004, while the Parallel Model explained 20% of the variance, MSE = 0.1.

While the results are rather clear cut in favor of the Serial Model, several concerns present themselves: First, could the serial gap be explained by a very noisy parallel simulation? To test this, in **Experiment S1** we independently examined the perceptual noise surrounding object locations with a separate group of participants, and found that it is nowhere near the levels that would be relevant for such a claim. Second, perhaps there is something uniquely complicated about a situation involving realistic physics of objects falling under gravity, as opposed to the more simplified stimuli often used in MOT. To this objection, we would note that if anything, we should expect people to be better adjusted to the more ecological task of objects colliding and falling under gravity than the not-frequently-encountered task of objects moving in free-form, and so people should be more likely to exhibit *greater* capacity in ecologically valid tasks. Still, to examine this empirically, in **Experiment S2** we tested participants in a simplified task in which objects moved more like hockey pucks on a smooth surface as seen from a top-down view, without collision, and not under gravity (similar to many MOT tasks). We found the same pattern of results as in Experiment 1b: a large delay between the responses, in line with a Serial Model. A third concern is that people may actually be able to carry out a parallel simulation in principle, but simply choose not to in practice, because such a simulation is more effortful than a serial simulation. This concern faces several in-principle difficulties: participants are also presumably motivated by opportunity costs to finish the task quickly, so why not finish it faster through parallel simulation? And why would the total effort of serial simulation over longer periods be less than that of parallel simulation over shorter periods? Beyond such theoretical issues, we empirically tested the motivation concern in **Experiment S3**, which was similar to Experiment 1b except that we informed participants they would be paid a bonus to the degree to which they were close to the ground truth timing. We found that a motivation manipulation had no effect, and replicated instead the findings of Experiment 1b. Fourth, it could be that capacity limits merely reflect some disruption to tracking from the fact the balls froze mid-air instead of disappearing in a more ecological way (see 24). To rule this out, in **Experiment S4** we tested how well participants estimate the impact time of balls that do not freeze but simply gradually disappear behind an occluder. This naturalistic situation replicated the serial pattern found in Experiment 1b. A fifth concern may be that averaging across individuals hides important variation, such that some people are able to simulate two or more objects in parallel. However, an **individual differences analysis** shows this is not the case. The full rationale and methods of these additional experiments and analyses are detailed in the Supplementary Information, but to summarize briefly, our conclusion from them and the results of Experiment 1b is that tracking imagined objects via mental simulation is limited to as little as a single object.

Experiment 2: Tracking two objects in perception

The results of Experiment 1b suggest an extreme capacity limit in the imagination, such that people only simulated the motion of a single object at a time. However, a major objection is that the bottleneck exists due to a serial *response* process, instead of in the simulation process. Notably, the requirements of response *selection* were deliberately minimized: the task involved a constant response mapping, responses were congruent with the side in which each ball appeared, separate hands were used for the two response keys, and participants pressed each key once on each trial. Also, if the bottleneck was such that simulation happened in parallel, but motor-delay caused a constant delay in execution, then we would expect to see a *constant* additive factor that does not depend on the true impact time of the objects, which contrasts with our findings (for further evidence from individual differences, see the Supplementary Information).

Still, to more directly test the possibility that the serial bottleneck was created by response *execution* rather than mental simulation, we conducted Experiment 2. The response requirements of this experiment were identical to Experiment 1b, but the same scenes now played all the way through, meaning participants saw the balls actually hit the ground, without the need to imagine their future paths (see Fig. 4, left). If the serial pattern of Experiment 1b reflects any response-related factor, the results of Experiment 2 should replicate it. But, if the serial pattern is specifically due to the need to simulate the future trajectory of objects, Experiment 2 should be closer to the Parallel Model's predictions.

We found that participants performed well overall, with a linear modulation of subjective impact time by true impact time, F(1.4, 49.04) = 570.03, $p < 0.001$, partial $\eta^2 = 0.94$; linear trend: t(105) = 41.3, $p < 0.001$. As can be seen in Figure 4 (right), instead of replicating the serial pattern of Experiment 1b, the results of Experiment 2 revealed a much smaller response delay. The second responses were slower, F(1, 35) = 60.15, $p < 0.001$, partial $\eta^2 = 0.63$, in a way that now interacted with true impact time, F(2.36, 82.62) = 31.74, $p < 0.001$, partial $\eta^2 = 0.47$, due to a larger effect for smaller impact times. Critically, the effect of response order was much smaller than in Experiment 1b, F(1, 70) = 74.65, $p < 0.001$, partial $\eta^2 = 0.52$. As can be seen also in the Parallel Model's predictions, some effect of response order is always expected (by definition, the second response is slower than the first), but as Experiment 2 empirically shows, the effect is very small, averaging at 88 ms (CI for the intercept of the first response: [181, 239] ms, second response: [514, 678] ms). Model fits confirmed that the Parallel model was preferred for tracking in perception: the Parallel Model explained 97% of the variance in average responses, MSE = 0.001, while the Serial Model explained 47% of the variance, MSE = 0.02.

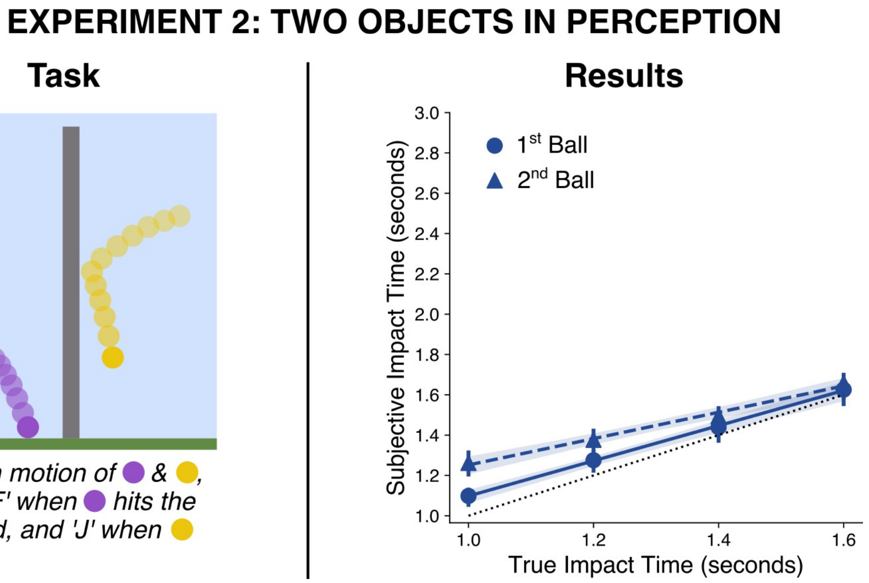


Figure 4: Task and results of Experiment 2, tracking two objects in perception. Circles and triangles indicate mean responses for different true impact time and response order ('1st ball' refers to the ball participants responded to first, and '2nd ball' to the ball they responded to second), error bars show SEM, solid line shows best linear fit, shaded area is 95% CI, and dotted line shows hypothetical perfect performance (where the subjective impact time equals the true impact time), as reference

.

We stress that we do not take the results of Experiment 2 to definitively reflect either parallel or serial tracking *perception*. While the results are more aligned with the Parallel Model, that model refers to mental simulation, and it is possible that in perception people are either carrying out the task in parallel, or through very rapid serial switching. Whether it is one or the other does not matter to our central point here: the results of Experiment 2 differed drastically from Experiment 1b, and show that the serial pattern of Experiment 1b are not due to a bottleneck in response requirements (which were identical in Experiment 2). Instead, the results likely reflect a specific serial constraint on simulating the paths of imagined objects.

Experiment 3: Two objects in the imagination with grouping

The finding that people mentally simulate a single object at a time is striking when considering the simplicity of the current task compared to real-world tasks, which regularly involve many objects that can move in complex paths. Mental simulation likely evolved to employ different hacks (16) that might overcome the serial bottleneck found here. One important strategy could leverage regularities in the environment, such as Gestalt cues, which improve performance in perceptual tracking (25). It seems reasonable to expect that if the motion paths of different objects are similar enough, the objects will be grouped in imagination, allowing their physics to be advanced in parallel. We tested this idea in Experiment 3, which used the same imagination task as in Experiment 1b, but with three important modifications (see Fig. 5): only hyperbole motion was used, the two balls always moved in the same direction (either to the left or right, instead of toward each other), and velocity was held constant. This meant that the visible motion sequence was identical for all items, to encourage participants to group the two balls in each scene. Because the true impact time was determined solely by a ball's initial height, the setup also created a greater opportunity for using heuristics instead of imagining the exact trajectory, which could be another way of overcoming the single item capacity limit.

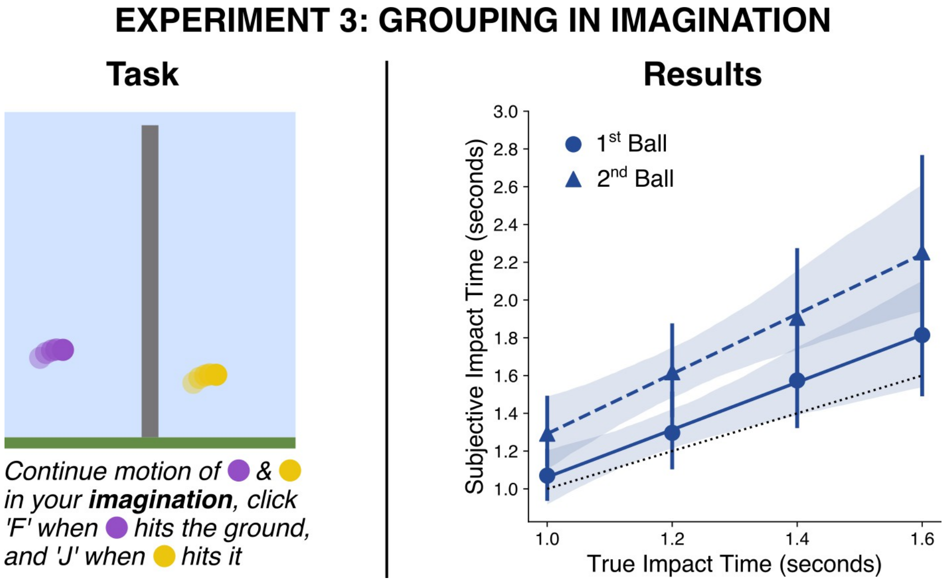


Figure 5: Task and results of Experiment 3, tracking two objects in the imagination with strong grouping cues. Circles and triangles indicate mean responses for different true impact time and response order ('1st ball' refers to the ball participants responded to first, and '2nd ball' to the ball they responded to second), error bars show SEM, solid line shows best linear fit, shaded area is 95% CI, and dotted line shows hypothetical perfect performance (where the subjective impact time equals the true impact time), as reference

. The large difference found for tracking independent objects in the imagination was diminished but not abolished.

Participants performed the task reasonably well, and the subjective impact times were linearly modulated by true impact time, F(1.05, 36.8) = 36.84, $p < 0.001$, partial $\eta^2 = 0.51$; linear trend: t(105) = 10.51, $p < 0.001$. As can be seen in Figure 5 (right), the second responses were still much slower than the first, F(1, 35) = 62.65,

$p < 0.001$, partial $\eta^2 = 0.64$, in a way that interacted with true impact time, F(2.17, 75.84) = 4.1, p = 0.02, partial $\eta^2 = 0.1$, this time because of a larger difference for larger impact times. The average difference between the first and second responses was 328 ms (CI for the intercept of the first response: [-461, 47] ms, second response: [-670, 28] ms), which was smaller than in Experiment 1b, F(1, 70) = 17.13, $p < 0.001$, partial $\eta^2 = 0.2$, but larger than in Experiment 2, F(1, 70) = 31.17, $p < 0.001$, partial $\eta^2 = 0.31$. Accordingly, we found an intermediate pattern based on the fit between participants' data and our computation models: the Serial Model explained 69% of the variance in responses, MSE = 0.04, and the Parallel Model explained 86% of the variance, MSE = 0.02. The results suggest that grouping could relax the single item bottleneck of imagination tracking somewhat, but not eliminate seriality completely, even with identical motion sequences and an opportunity to use heuristics.

Discussion

Research spanning decades has demonstrated that people have signature capacity limits when tracking visible objects. Here, we examined capacity limits when objects were moving in the imagination. We found that the mind's eye can only track a single object at a time. More specifically, we found that people could reasonably unfold the trajectory of a single object in the imagination (Experiment 1a), but that the addition of just one independent object substantially altered their responses (Experiment 1b), in line with the predictions of serial mental simulation. This Serial Model suggests people first had to mentally advance one object up to some point, before going back and advancing the second object. We did not observe the capacity bottleneck when people tracked two objects in perception instead of imagination (Experiment 2), further cementing the notion that the capacity limit is in mental simulation, not the motor response or other limits further downstream. Additional experiments, models, and analyses (see the Supplementary Information) showed that the serial gap is not the result of noisy parallel simulation, lack of motivation, the need to simulate realistic physics, or disruption due to the unnatural freezing we used, and also that the limitations hold at the individual participant level. Notably, the stable serial pattern emerged despite of the well-known difficulty of observing serial costs in performance (the difficulty of teasing apart serial and parallel patterns applies when a seemingly parallel pattern could be interpreted as very rapid serial switching, but that is not the case here). When we added strong grouping cues to the trajectory of the objects, we found that the difference between the first and second response shrank, but was not fully eliminated (Experiment 3). Taken together, our results suggest that mentally simulating the movement of objects is a serial process.

The finding that people are able to track only a single object at a time in their imagination is surprising, seeing as people can usually track a handful of items in direct perception (though the exact number is affected by various factors, such as object speed or spacing; e.g., 2; 8; 1; 5). If seeing things in the mind's eye is supposed to be akin to seeing with one's real eyes (26), our findings suggest it isn't so. However, while researchers do use the term 'track' to include the following of hidden objects behind occluders in perceptual tasks, it may be that this is an over-loaded term. 'Tracking' in the imagination (or through occlusion) may be quite different than direct perceptual tracking, as it is the mind itself that is moving the objects, rather than keeping on top of objects that are being moved by external forces. Such a distinction aligns well with two lines of work in attention and working memory. First, people's ability to extrapolate motion in perceptual tracking was recently suggested to have a capacity limit of only one object (6), perhaps due to challenges of physical simulation (27). Second, updating active representations was argued to depend on sequentially loading objects, one at a time, into the 'focus of attention' (28). So, it may be that calculating an object's future motion (whether in direct perception as in MOT, or in imagination as in Imagined Objects Tracking) requires constantly updating the object's representation in working memory, and that relies on a serial process (for additional connections between physical simulation and working memory, see 29). This is further strengthened by the present finding that the serial pattern can be observed not only for items that freeze mid-motion, but for items that undergo natural occlusion (Experiment S4). The capacity limits we found in the imagination should then be taken to refer to the simulation part that moves objects forward, rather than to a later process that re-processes the imagined scene. Also, we did not control for eye movements, and it is possible that people followed imagined trajectories with their eyes, which contributed to the single-object bottleneck, although this only raises the question of why people could not shift their eyes to track both imagined objects (as they do in perception). This is not a limitation of the studies, but rather is in line with how people may carry out physical predictions (see, e.g., 30; 31; 32), and is an interesting topic for future research.

Independent of capacity limits in perception, another reason why our findings are surprising is that they contrast with subjective intuitions about internal scenes. Many people report being able to imagine vividly dynamic mental scenes, and a single-object capacity in simulation doesn't align with that. Why then does it subjectively seem like we can imagine vividly moving dynamic scenes? This is similar to the apparent conflict between our intuition and other capacity limits – for example, in our everyday life, we do not feel like we only have access to a tiny subset of all of the perceptual input, yet decades of working memory research have shown that this is indeed the case, and under naturalistic conditions we simply rely on other mechanisms for compensation, such as long-term memory or scene scanning with saccades (for a review of similar ideas, see 33). Aside from general arguments regarding the unfaithful nature of introspection (e.g., 34), our third study showed that grouping does ameliorate the serial effect (though it doesn't negate it). Our main effect relied on intentionally creating scenes in which mental objects move independently of one another, but this may not be a typical case. It is likely that many dynamic mental scenes (perhaps also those previously used in intuitive physics research; e.g., 15; 35) rely on strong grouping and hierarchical organization, such that the serial process need only update a hyper-parameter that controls the motion of several objects at the same time. One such example would be mentally simulating the distance between items, instead of the location of each item separately, an interesting idea that can be the target of future work.

Our studies focused on non-interacting objects to keep the findings clear and simple, but objects in the mind can interact. A simple case-study of minimal interaction is two objects moving along a plane at various speeds and angles, possibly about to collide. In such a case it seems unlikely that people use a serial updating process that fully moves first one object, then another, as no collisions would occur. In such a case, perhaps people move forward one object for a limited number of steps S, then switch to another object, and cycle back again. As detailed in the Supplementary Information, such 'interleaved' serial simulation models did not explain the data in the present studies, but they may be relevant for interaction/collision situations, possibly with a dynamically set S. We plan to pursue such cases and models in future studies.

Another intriguing direction for future research concerns the information that people do manage to simulate. Specifically, it is unclear whether people imagine the objects along with all of their features, or are closer to a computerized physics engine that handles only trajectories. The current results cannot offer an answer to this question, and it is independent from the issue of the capacity limit of the simulation process. Yet, past findings from MOT do point to a differential status of spatiotemporal information and surface features at least in perceptual tracking. An extreme manifestation of this is that while featural information can definitely aid tracking by allowing for more efficient deployment of attentional resources (e.g., 36), when the tracked objects change their features, people might entirely miss this (e.g., 37). On the other hand, MOT findings suggest that people manage to rely on featural information for grouping (e.g., 25), and so it would be interesting to test whether the single object capacity limit in mental simulation might be relaxed not only by physically-relevant information (as the identical motion paths used in Experiment 3), but also by presenting the objects in the same color.

The capacity limits we found hold independently of the specific cognitive computations one assumes people use to advance objects in the mind's eye. That said, we do adhere to a mental simulation approach to intuitive physics (e.g., 14), and our computational models did assume that people track objects in the imagination by mentally advancing them in step-by-step. This view contrasts with research that argues that humans do not rely on mental simulation for intuitive physical judgments, and which often points to people's systematic mistakes and deviations from ground-truth physics as evidence against simulation. While these two views are often portrayed in opposition, we see the current work as another brick in the bridge between the rich literature on errors in physical judgments (e.g., 20; 38) and the mental game engine framework. It is part of a general approach that uses game engines as inspiration for an overall mental simulation account, but also draws on the shortcuts and workaround used in such engines to save on time, memory, and overall computation (16). Such an approach has found evidence for people's use of systematic approximations in the representations of bodies themselves (39), as well as people's use of 'partial simulation', in which they do not mentally simulate parts of the scene that are deemed irrelevant (21). Our present work shows another central way in which mental physics parts ways with real physics, while still being overall consistent with a mental simulation account.

We set out to examine the capacity limits of the imagination. We found that even in a simple situation the answer to 'how many objects can the mind's eye keep track of at once?' is 'approximately one'. This might feel like discovering you've been tricked. Like realizing that who you took to be a fantastic juggler is really only tossing and bouncing a single ball. Still, knowing the trick makes you appreciate the act in a different way: It's poor juggling, but it's a great trick.

Methods

Materials, data, code, and pre-registration protocols for all experiments, are available in the following Open Science Framework repository: `https://osf.io/wzt98/`.

Participants

Research was approved by the Harvard University Ethics Committee (protocol IRB19-1861). All participants provided informed consent. Participants were recruited online (40) via Prolific (`https://www.prolific.com`). They were paid $1.6, and the median time to complete the studies ranged between 5.5 and 6.5 minutes. Participation was restricted to English-speaking US-based participants, with an approval rate of at least 95%, who did not perform any of the other tasks in the study (including pilot studies, see below).

Given that Imagined Objects Tracking is a novel task, we ran pilot studies (with the same tasks described in the pre-registration; data available at the OSF) to determine the necessary sample size for both within- and between-subjects comparisons. The smallest effect size found (an interaction of a within-subjects effect and experiment) was partial $\eta^2 = 0.19$, which requires N = 18 in each experiment for 95% power with $\alpha = 0.05$ (calculated using G*Power 3, 41). As a conservative estimate, we decided to double this number in the full study. In the case of participants failing the comprehension questions and being excluded, we recruited additional participants to reach 36. All decisions of screening and re-recruitment were based on pre-registered criteria, and took place without analyzing the data itself.

Participants were excluded from all analyses if they gave an incorrect answer to at least one of the pre-task quiz questions, or (in experiments with 2 entities) if less than 75% of their trials included two unique responses (i.e., two different response keys). To ensure N = 36 participants in the final sample of each experiment, this required recruiting N = 47, N = 71, N = 53, and N = 56 participants in Experiment 1a, 1b, 2, and 3, respectively. This was a comparable rate to similar past studies of intuitive physics conducted online (e.g., 21; 19). The pre-task quiz and unique responses threshold were the only criteria used for excluding participants, intentionally focusing only on task comprehension rather than task performance. The final sample in Experiment 1a included 16 people who identified as female, 19 as male, and one who preferred not to state (mean age 34.4); Experiment 1b included 18 people who identified as female, 18 as male (mean age 38.6); Experiment 2 included 20 people who identified as female, 16 as male (mean age 40.3); Experiment 3 included 22 people who identified as female, 13 as male, and one who preferred not to state (mean age 37.0).

Stimuli and Procedure

We used an animated dynamic prediction task, similar to tasks previously used to study intuitive physical reasoning (e.g. 21; 19). Participants in the Imagined Objects Tracking task continue the trajectory of objects in their mind's eye, and a measure of the tracked motion is compared with the ground truth. In the case of the present experiments, the measure is the time in which an event happened in imagination (a collision with the ground), and the ground truth is extracted from the physics engine used to create stimuli. Demos of the tasks are available online: https://jatos.mindprobe.eu/publix/Z8AtkMP8NZt (Experiment 1a), https://jatos.mindprobe.eu/publix/5iB7OvSVmHX (Experiment 1b), https://jatos.mindprobe.eu/publix/1ygkkWoPJZm (Experiment 2), and https://jatos.mindprobe.eu/publix/4vh34OQB263(Experiment 3).

In all experiments, participants watched short 2D animations, created in the physics engine Pymunk, that used the same simple setting: A green rectangle at the bottom represented the ground, a narrow upright gray rectangle at the horizontal mid-line represented a wall, and a light blue rectangle acted as background. Additionally, each scene included 1 or 2 balls, rendered as yellow or purple disks. For scenes with 2 balls, one was always to the left of the wall and the other was to the right, and the balls differed in color. Scenes started with each ball having some initial height and velocity, after which the balls moved according to simulated physics. We manipulated the initial height and velocity to produce different trajectories that varied in their paths and the time it took a ball to hit the ground, which was either 1.0, 1.2, 1.4, or 1.6 seconds from the start of the animation.

The task in all experiments was to indicate when a ball touches the ground. Response keys were spatially mapped to avoid confusion: 'F' for balls left of the wall, and 'J' for balls right of the wall. No feedback was given following button presses. Each combination of true impact time and ball movement type was presented 8 times (4 on each side, randomized order), for a total of 64 experimental trials, presented in 2 blocks with a self-timed break between them.

Prior to the test trials, participants went through a pre-task phase, including instructions, practice trials, and a multiple-choice quiz. Each question in the quiz focused on a different aspect of the task (the goal, when animations terminate, how to be most accurate, and response mapping). If a participant failed to respond correctly to any of the 4 quiz questions they were removed from further analysis. We next provide details specific to the setup of each Experiment.

Experiments 1a and 1b. Animations in the experimental phase paused after 0.5 sec (well before either ball touched the ground). Participants were asked to continue the animation in their mind's eye, and indicate when the balls in their imagination hit the ground. The starting height and velocity of the balls were chosen such that neither variable on its own determined the time it took for the ball to reach the ground, to discourage the use of heuristics. During practice, participants completed 4 trials with animations that ran all the way through (showing the impact of the ball with the ground), in which they were asked to press a button when they saw the ball touching the ground. This was followed by 4 trials with animations that paused early (not showing the impact), as in the actual experiment. Animations in Experiment 1a showed a single ball (see Fig. 2, left), and animations in Experiment 1b showed 2 balls (see Fig. 3, top left)). In two-entity animations, one ball moved in a hyperbole up and to the center, without hitting the wall (from shortest to longest true impact time, these balls started either 100, 140, 60, or 180 pixels above the ground, and their vertical velocity was 95, 110, 216, or 180 pixels per second; their initial distance from the wall was 185 pixels, and their horizontal velocity was 100 pixels per second), and the other ball moved down and towards the wall on a sure collision path with it, but with the moment of collision occurring after the animations paused (from shortest to longest true impact time, these balls started either 280, 480, 400, or 440 pixels above the ground, and their vertical velocity was 190, 200, 60, or 20 pixels per second; their initial distance from the wall was 185 pixels, and their horizontal velocity was 280 pixels per second). The color of the balls was matched to movement type within participants, but randomized between participants. Each movement type was counterbalanced to appear on each side of the wall on half of the trials. Single ball animations were created by removing one of the balls from the 2-balls scenes. In Experiment 1a, each block included only balls either left or right of the wall, with the order counterbalanced across participants.

Experiment 2. The stimuli were identical to Experiment 1b (2 entities), except that the animations continued until participants gave 2 responses, including the moment in which the balls touched the ground, and up to 4 seconds (see Fig. 4, left). So, rather than continue the motion of objects in their imagination, participants were asked to simply click on the appropriate button when they saw the relevant ball touch the ground. Accordingly, the practice phase included 4 full-length animations, without imagination trials.

Experiment 3. The task was identical to Experiment 1b (2 entities, animations pause, participants continue the motion in their imagination). However, unlike Exp 1b, all balls moved in a hyperbole motion, in the same direction (left or right), and velocity was kept constant (see Fig. 5, left). This was designed to create strong motion grouping cues. The only variable that led to different true impact times was the starting height of each ball. The color mapping (yellow/purple ball shown on left/right of wall) was randomized between participants, but constant within participants.

Analysis

Responses were aggregated across movement type, color, and location relative to the wall. In Experiment 3, all motion paths were hyperbolic, and trials were aggregated across movement direction (towards or away from the center). Individual trials were rejected from further analysis if they did not include two unique responses, as this prevents mapping each response to a specific ball. This rejection was not applied to Experiment 1a, as it involved only one ball in each trial. Trials were also rejected if responses were farther than 3 SDs from a participant's mean. Taken together, these criteria resulted in a rejection of less than a single trial on average in all experiments: 0.3, 0.3, 0.2, and 0.5 trials on average in Experiment 1a, 1b, 2, and 3, respectively (note that the values reflect the number of rejected trials, not the percentage of rejected trials, which was 0.5%, 0.5%, 0.3%, and 0.8%, respectively, all lower than 1% of rejected trials).

Statistical Tests. In all experiments, we analyzed Subjective Impact Time using a within-subject Analysis of Variance (ANOVA) with True Impact Time (1.0, 1.2, 1.4, or 1.6 sec; extracted from the physics engine) as a factor. In experiments that included 2 entities, we added Response Order (first vs. second key press) as a factor. We followed the ANOVAs with a polynomial contrasts analysis, to test the linear trend of the True Impact Time factor. As a measure of the effect of Response Order in experiments with 2 entities, we used 1,000 bootstrap samples to estimate the 95% confidence interval (CI) on the intercept of linear fits, separately for the first and second responses. Our pre-registered predictions were to find (1) a linear trend for all experiments, showing that overall people are sensitive to ground truth physics; (2) a large delay between responses in Experiment 1b, in line with a Serial model; (3) a reduced effect in Experiment 2; and (4) an intermediate effect in Experiment 3. Because the effect of response order was expected to be significant even for Experiment 2 (given that the second response is by definition slower), we additionally compared the effect of Response Order across experiments, using ANOVAs with Response Order as a within-subjects factor, and Experiment as a between-subjects factor, and predicted significant interactions. The Supplementary Informationfurther reports a post-hoc analysis focused on the variation in response times. Violations of sphericity were handled via Greenhouse-Geisser corrections (42). All tests are two-tailed.

Parallel vs. Serial Mental Simulation Models. We created two mental simulation models: Serial and Parallel. Both models relied on the same physics engine that generated the stimuli to simulate the balls, starting from the animation's end point and until both balls collide with the ground. The models differed in how they advanced the objects (see Fig. 1). The Parallel Model moves both balls simultaneously. The Serial Model first picks one ball, advances its state until collision with the ground, then repeats this process for the second ball. More formally, taking a scene to be a tuple of objects o_t^j at time t, and each object to be a list maintaining the properties of the object at time t, and Φ to be the transition function that updates the properties of objects according to physics, we have for the Parallel Model:

$$
\begin{gathered}
t = 0 : [o_0^1, o_0^2] = G + \xi, \\
t > 0 : [o_{t+1}^1, o_{t+1}^2] = [\Phi(o_t^1), \Phi(o_t^2)],
\end{gathered}
$$

where G is the ground-truth state of the objects as handed by perception, and ξ is the perceptual noise. Following standard practice in modeling intuitive physics with mental simulation (e.g., 14) we assume this is a two-dimensional Gaussian with mean $\mu = (0, 0)$ and a symmetrical standard deviation $SD = (\sigma^j, \sigma^j)$ for each object j. We estimated an upper level for perceptual noise in an independent experiment and used this value as our noise level, but importantly, our results are extremely robust both below and above this chosen perceptual uncertainty setting, including both no-noise situations, and far greater noise levels (for the full details, see the Supplementary Information).

For the Serial Model, we have:

$$
\begin{gathered}
t = 0 : [o_0^1, o_0^2] = G + \xi, \\
t > 0 \wedge t < C : [o_{t+1}^1, o_{t+1}^2] = [\Phi(o_t^1), o_t^2], \\
t > C : [o_{t+1}^1, o_{t+1}^2] = [o_t^1, \Phi(o_t^2)],
\end{gathered}
$$

where the choice of simulating object o^1 first is arbitrary, and C is the time at which object o^1 collides with the ground. While our main analysis takes the choice of which object to simulate first to be random, we do expect that people are biased in this selection, and indeed we found evidence that people use simple imperfect cues in this selection (see the Supplementary Information).

Model Fitting. Because the models include perceptual uncertainty, we sampled 20 starting-states for each model, and averaged the results across runs. In addition to the perceptual uncertainty parameter that was estimated through independent participant data (Experiment S1), we assume that the model response can be fit to the human response up to a simple linear transformation, meaning $Human\ Subjective\ Impact\ Time = a \cdot (Model\ Predicted\ Impact\ Time) + b$. We fit the slope and intercept of this linear transformation for each model separately, using the response data of the relevant experiments that involve 2 objects. To compare model performance, we calculated each model's explained variance, as well as the resulting MSE. Model parameter fits were done for the mean responses of all participants. In the Supplementary Information, we additionally present model fits for individual data (still fitting overall a and b). All of these different analyses agree with the results of the analysis we present in the main text.

Acknowledgments

We thank members of the CoCoDev labs at Harvard, and the CoCoSci lab at MIT, for their insightful comments. This work is supported by the Center for Brains, Minds, and Machines (HB, TDU), the Defense Advanced Research Projects Agency (DARPA) Machine Common Sense Program (HB, TDU), and the Jacobs Foundation (TDU).

Competing interests statement

The authors declare no competing interests.

Data Availability

Data for all experiments (including pilot studies) can be found at `https://osf.io/wzt98/`.

Code Availability

The code for the stimuli, models, and analysis can be found at `https://osf.io/wzt98/`.

References

[1] Z. W. Pylyshyn and R. W. Storm, "Tracking multiple independent targets: Evidence for a parallel tracking mechanism*," *Spatial Vision*, vol. 3, no. 3, pp. 179–197, 1988.

[2] G. A. Alvarez and S. L. Franconeri, "How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism," *Journal of Vision*, vol. 7, p. 14, Oct. 2007.

[3] Z. W. Pylyshyn, "Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities," *Visual Cognition*, vol. 11, pp. 801–822, Oct. 2004.

[4] J. M. Scimeca and S. L. Franconeri, "Selecting and tracking multiple objects," *WIREs Cognitive Science*, vol. 6, pp. 109–118, Mar. 2015.

[5] B. J. Scholl, "What Have We Learned about Attention from Multiple-Object Tracking (and Vice Versa)?," in *Computation, Cognition, and Pylyshyn* (D. Dedrick and L. Trick, eds.), pp. 49–78, The MIT Press, June 2009.

[6] A. Holcombe, *Attending to Moving Objects.* Cambridge University Press, 1 ed., Feb. 2023.

[7] C. S. Feria, "Speed has an effect on multiple-object tracking independently of the number of close encounters between targets and distractors," *Attention, Perception, & Psychophysics*, vol. 75, pp. 53–67, Jan. 2013.

[8] S. Franconeri, S. Jonathan, and J. Scimeca, "Tracking Multiple Objects Is Limited Only by Object Spacing, Not by Speed, Time, or Capacity," *Psychological Science*, vol. 21, pp. 920–925, July 2010.

[9] A. Lovett, W. Bridewell, and P. Bello, "Selection enables enhancement: An integrated model of object tracking," *Journal of Vision*, vol. 19, p. 23, Dec. 2019.

[10] S. L. Franconeri, Z. W. Pylyshyn, and B. J. Scholl, "A simple proximity heuristic allows tracking of multiple objects through occlusion," *Attention, Perception, & Psychophysics*, vol. 74, pp. 691–702, May 2012.

[11] B. Keane and Z. Pylyshyn, "Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function," *Cognitive Psychology*, vol. 52, pp. 346–368, June 2006.

[12] L. Iordanescu, M. Grabowecky, and S. Suzuki, "Demand-based dynamic distribution of attention and monitoring of velocities during multiple-object tracking," *Journal of Vision*, vol. 9, pp. 1–1, Apr. 2009.

[13] A. Ahuja and D. L. Sheinberg, "Behavioral and oculomotor evidence for visual simulation of object movement," *Journal of Vision*, vol. 19, p. 13, June 2019.

[14] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 18327–18332, Nov. 2013.

[15] T. Gerstenberg, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum, "A counterfactual simulation model of causal judgments for physical events.," *Psychological Review*, vol. 128, pp. 936–975, Oct. 2021.

[16] T. D. Ullman, E. Spelke, P. Battaglia, and J. B. Tenenbaum, "Mind Games: Game Engines as an Architecture for Intuitive Physics," *Trends in Cognitive Sciences*, vol. 21, pp. 649–665, Sept. 2017.

[17] T. Gerstenberg, M. F. Peterson, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum, "Eye-Tracking Causality," *Psychological Science*, vol. 28, pp. 1731–1744, Dec. 2017.

[18] J. B. Hamrick, P. W. Battaglia, T. L. Griffiths, and J. B. Tenenbaum, "Inferring mass in complex scenes by mental simulation," *Cognition*, vol. 157, pp. 61–76, Dec. 2016.

[19] E. Ludwin-Peery, N. R. Bramley, E. Davis, and T. M. Gureckis, "Broken Physics: A Conjunction-Fallacy Effect in Intuitive Physical Reasoning," *Psychological Science*, vol. 31, pp. 1602–1611, Dec. 2020.

[20] E. Ludwin-Peery, N. R. Bramley, E. Davis, and T. M. Gureckis, "Limits on simulation approaches in intuitive physics," *Cognitive Psychology*, vol. 127, p. 101396, June 2021.

[21] I. Bass, K. A. Smith, E. Bonawitz, and T. D. Ullman, "Partial mental simulation explains fallacies in physical reasoning," *Cognitive Neuropsychology*, vol. 38, pp. 413–424, Nov. 2021.

[22] R. Keogh and J. Pearson, "The perceptual and phenomenal capacity of mental imagery," *Cognition*, vol. 162, pp. 124–132, May 2017.

[23] C. R. Ceja and S. L. Franconeri, "Difficulty limits of visual mental imagery," *Cognition*, vol. 236, p. 105436, July 2023.

[24] B. J. Scholl and Z. W. Pylyshyn, "Tracking Multiple Items Through Occlusion: Clues to Visual Objecthood," *Cognitive Psychology*, vol. 38, pp. 259–290, Mar. 1999.

[25] G. Erlikhman, B. P. Keane, E. Mettler, T. S. Horowitz, and P. J. Kellman, "Automatic feature-based grouping during multiple object tracking.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, pp. 1625–1637, Dec. 2013.

[26] S. M. Kosslyn, W. L. Thompson, and G. Ganis, *The case for mental imagery*. Oxford university press, 2006.

[27] J. S.-H. Lau and T. F. Brady, "Noisy perceptual expectations: Multiple object tracking benefits when objects obey features of realistic physics.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 46, pp. 1280–1300, Nov. 2020.

[28] K. Oberauer, "Access to information in working memory: Exploring the focus of attention.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 28, no. 3, pp. 411–421, 2002.

[29] H. Balaban, K. Smith, J. Tenenbaum, and T. D. Ullman, "Electrophysiology reveals that intuitive physics guides visual tracking and working memory," preprint, PsyArXiv, May 2023.

[30] M. J. Spivey and J. J. Geng, "Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects," *Psychological Research*, vol. 65, pp. 235–241, Nov. 2001.

[31] A. Pathak, S. Patel, A. Karlinsky, S. Taravati, and T. N. Welsh, "The "eye" in imagination: The role of eye movements in a reciprocal aiming task," *Behavioural Brain Research*, vol. 441, p. 114261, Mar. 2023.

[32] K. Krasich, K. O'Neill, and F. De Brigard, "Looking at Mental Images: Eye-Tracking Mental Simulation During Retrospective Causal Judgment," *Cognitive Science*, vol. 48, p. e13426, Mar. 2024.

[33] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behavioral and Brain Sciences*, vol. 24, pp. 87–114, Feb. 2001. Publisher: Cambridge University Press.

[34] D. C. Dennett, *Consciousness explained.* Penguin UK, 1993.

[35] D. M. Bear, E. Wang, D. Mrowca, F. J. Binder, H.-Y. F. Tung, R. T. Pramod, C. Holdaway, S. Tao, K. Smith, F.-Y. Sun, L. Fei-Fei, N. Kanwisher, J. B. Tenenbaum, D. L. K. Yamins, and J. E. Fan, "Physion: Evaluating Physical Prediction from Vision in Humans and Machines," 2021. Publisher: arXiv Version Number: 3.

[36] T. Makovski and Y. V. Jiang, "Feature binding in attentive tracking of distinct objects," *Visual Cognition*, vol. 17, pp. 180–194, Jan. 2009.

[37] B. Bahrami, "Object property encoding and change blindness in multiple object tracking," *Visual Cognition*, vol. 10, no. 8, pp. 949–963, 2003.

[38] M. McCloskey, A. Caramazza, and B. Green, "Curvilinear Motion in the Absence of External Forces: Naïve Beliefs About the Motion of Objects," *Science*, vol. 210, pp. 1139–1141, Dec. 1980.

[39] Y. Li, Y. Wang, T. Boger, K. A. Smith, S. J. Gershman, and T. D. Ullman, "An approximate representation of objects underlies physical reasoning.," *Journal of Experimental Psychology: General*, June 2023.

[40] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, vol. 70, pp. 153–163, May 2017.

[41] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods*, vol. 39, pp. 175–191, May 2007.

[42] S. W. Greenhouse and S. Geisser, "On methods in the analysis of profile data," *Psychometrika*, vol. 24, pp. 95–112, June 1959.

Supplementary Information for *The capacity limits of mental simulation*

The materials and data for all supplementary studies and analyses can be found in the following Open Science Framework repository: `https://osf.io/wzt98/`.

Serial vs. Parallel Simulation at the Individual Level

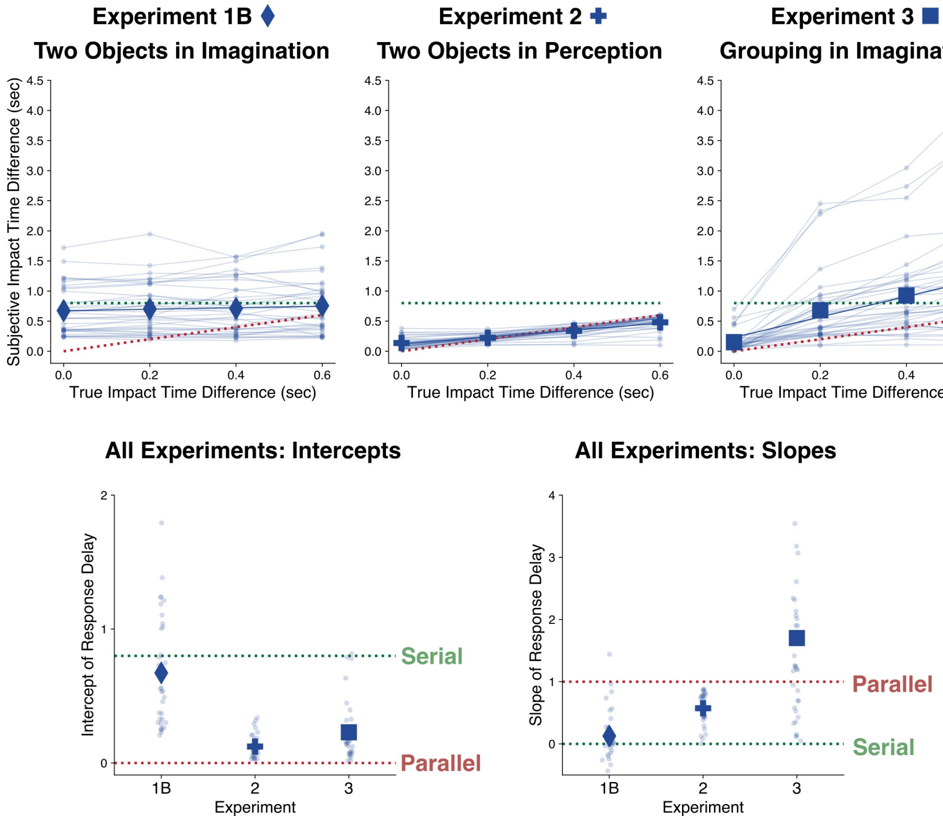
Across the main experiments with two objects, we also examined individual performance. Specifically, the models make different predictions regarding the delay in response between the balls, as a function of the true impact time difference between them. In a Parallel simulation, the delay should always closely follow the true difference, because both balls are imagined simultaneously, and when the first one hits the ground, participants only have to simulate what is left of the second ball's motion, which matches the difference in impact time between the balls. Therefore, the parallel prediction is a slope of 1 and an intercept of 0. Conversely, a completely Serial simulation predicts a constant delay on average, i.e., a slope close to 0, because after the first ball hits the ground, the second simulation has to start from the beginning, meaning that the delay in simulation should only reflect the unseen duration of of the second ball's motion (which doesn't depend on the first ball). With the same logic, the intercept should be around 800 ms, which is the average duration of the unseen motion of a ball (note that the order in which the balls actually hit the ground is unknown to the model, and similarly for participants). Importantly, note that the average delay of 800ms reflects an averaging of *differential* effects, rather than a fixed constant response across trials. If the serial gap was due to a parallel simulation followed by a fixed gap in response due to a down-the-line bottleneck (such as motor delay), we should expect a slope of 1 but an intercept greater than 0.

The results of the individual participants' data analyses are shown in Fig. S1. For easier comparison across experiments, the raw results (Fig. S1, top) are complemented by distributions of individual intercept and slope values from a linear fit to the raw data (Fig. S1, bottom). In short, the individual-level analysis corroborates the findings in the main text.

In Experiment 1b, there is a larger-than-zero delay for all of the participants, averaging at 711 ms. This is true even when the difference in true impact time is 0, meaning that the balls would have hit the ground at exactly the same time if the scene would unfold all the way through. Additionally, the flat delay pattern, meaning no modulation by the difference between the balls, is highly robust at the individual level. Both of these findings provide further support for a serial simulation, and demonstrate that the reported effects are not the result of averaging over individual patterns that do not resemble the mean.

In Experiment 2, when objects are visible in perception, the intercept was close to 0 for many participants (mean: 123 ms). Additionally, individual slopes are also much closer to 1 than in Experiment 1b. This pattern is in line with parallel processing, where participants followed the two balls together.

In Experiment 3, where the mind's eye had to track objects that moved in identical ways, individual results show how participants could take advantage of the strong grouping cues, or the chance to employ heuristics, and track the balls in their imagination in a way that resembled the Parallel Model's predictions. This was true both for the intercept (reflecting the delay for balls with the same true impact time) which was close to 0 for many participants (mean intercept was 230 ms), and for the slope, which was close to 1 for many participants.



Experiment 1B ♦ Two Objects in Imagination

Experiment 2 ✚ Two Objects in Perception

Experiment 3 ■ Grouping in Imagination

All Experiments: Intercepts

All Experiments: Slopes

Figure S1: Individual response analysis. Response delay is shown as a function of the true difference in impact time between the two balls. Top: raw individual (small markers) and average (large markers) results in Experiments 1b, 2, and 3. Bottom: distributions of the intercept (left) and slope (right) values across experiments. Dotted lines show the predictions of the Parallel and Serial Models.

Interestingly, there was great variability in leveraging the grouping cues and/or heuristics, which highlights an interesting direction for future studies on how people overcome capacity limits in mental simulation.

Response Variation Analysis

Our main, pre-registered, analysis focused on average response time as a measure of subjective impact time. Another interesting angle to examine the data from concerns the variation in participants' responses. Specifically, if simulation indeed happens serially, we can expect the second responses to have more variation than the first responses, because by the time participants get to them more noise accumulates. This is similar to how longer simulations should also be noisier, and so with larger values of the true impact time, the responses should also have larger variation. Notably, not all of these effects can be attributed to the simulation, as some noise accumulates likely due to response and/or memory processes. Nevertheless, for completeness we also report post-hoc analysis of the individual level SD of responses across experiments.

In Experiment 1b, we found that response time SDs were larger (F(1, 35) = 9.11, p = 0.005, partial η^2 = 0.21) for the second response (95% CI: [374, 563]) than for the first response (95% CI: [286, 474]), and that SDs grew as a function of the true impact time (F(2.41, 84.37) = 4.93, p = 0.006, partial η^2 = 0.12), and these two factors did not interact (F < 1). This is in line with a serial simulation where noise accumulates not only with the duration of the simulations (as reflected by the true impact time), but also with the order in which these simulations are performed.

In Experiment 2, response time SDs were again larger (F(1, 35) = 35.12, p < 0.001, partial η^2 = 0.5) for the second response (95% CI: [99, 130]) than for the first one (95% CI: [69, 101]), but were not affected by the true impact time (F < 1), and these two factors again did not interact (F(2.72, 95.22) = 1.1, p = 0.35, partial η^2 = 0.03). Compared to Experiment 1b, the effect of true impact time was smaller (F(2.56, 179.63) = 4.94, p = 0.004, partial η^2 = 0.07), and that of response order was marginally smaller (F(1, 70) = 3.84, p = 0.054, partial η^2 = 0.05). Overall, the pattern is in line with the idea that here participants did not have to simulate the trajectories but only track them (hence the lack of an effect of the true impact time), and that at least some of the effect of response order on the variation is due to noise accumulated from the response process.

In Experiment 3, the results were in line with those of Experiment 1b, with larger SDs (F(1, 35) = 10.54, p = 0.003, partial η^2 = 0.23) for the second response (95% CI: [312, 497]) than for the first one (95% CI: [252, 437]), larger SDs with larger true impact times (F(2.07, 72.38) = 8.78, p < 0.001, partial η^2 = 0.2), and no interaction between these two factors (F < 1). Compared with Experiment 1b, the effect of response order was similar (F < 1), and the effect of true impact time was larger (F(2.41, 168.79) = 3.43, p = 0.027, partial η^2 = 0.05), mainly due to smaller SDs for shorter true impact times in Experiment 3.

Another interesting question is whether the difference in response times was also more varied as the true difference grew larger. Here, the Serial model's predictions are less clear cut: Because the second simulation starts from scratch, the model is not influenced by the difference in impact time itself (as can be seen in the previous section), but the different delays do represent different mixture of true impact times, and so with the larger delays the noise might be larger because of a larger proportion of long simulations. Therefore we report the results of this analysis with caution. We found marginal evidence for an effect of the true impact time difference on the SD of the response delay in Experiment 1b (F(1.67, 58.64) = 2.53, p = 0.097, partial η^2 = 0.07), no effect in Experiment 2 (F(2.24, 78.53) = 1.1, p = 0.35, partial η^2 = 0.03), and a strong effect in Experiment 3 (F(2.38, 83.29) = 21.66, p < 0.001, partial η^2 = 0.38).

To summarize, the results of the SD analyses are generally in line with the predictions of the Serial model. We stress that these tests were not pre-registered and we treat them as exploratory. Questions regarding differential variation in responses and the potential sources of noise that give rise to it remain an interesting direction for future research.

Experiment S1: Noise Estimation

Many current intuitive physics models based on mental simulation assume the presence of noise, due to perceptual or dynamic uncertainty (1). While the perceptual noise parameter was not our main focus, we note that the only way the Parallel Model can create any systematic difference between the response times is through the perceptual noise. So, it was important for us to examine whether a seemingly serial pattern of responses (a large difference between the first and second responses) is the result of a *very* noisy parallel simulation. To obtain an estimate of the perceptual noise in our main task, we ran a separate experiment using a modified paradigm with a group of participants independent from our main studies. We showed participants the same animations as those used in our main task, except that instead of freezing, the objects in the animation disappeared completely. Participants were asked to click on the location where the objects were last seen.

Method

We used the same stimuli as in Experiment 1b (two balls in the imagination) with a slight modification: after 500 ms, instead of freezing, the balls disappeared from view. Participants were asked to use their cursor to click on the location where the balls were right before they disappeared (for a demo, see `https://jatos.mindprobe.eu/publix/6m6OKyshQHx`).

Participants completed 64 trials, as in the main task of Experiment 1b. Also similar to Experiment 1b, participants went through a pre-task phase, which included instructions, practice trials, and a multiple-choice quiz. We collected responses from 25 Prolific participants, and excluded participants who failed to respond correctly to any of the 3 quiz questions (one participant), or did not provide at two mouse clicks on each trial (one participant).

Results

We were interested in the spatial spread of people's estimations, and examined the SD of their responses. For this, we aggregated the responses from all participants and all scenarios, and re-aligned all of the trials relative to one example scene. We then calculated the average location of responses for each ball, and the SD of responses relative to this average. We averaged across the two balls to reach one SD value, and found that the SD of responses was 30.6 pixels, corresponding to roughly 1.5x the radius of the balls (see Figure S2 for an illustration). This is the value we used in the main text as the SD of the perceptual noise distribution. It is likely that this overall uncertainty is mildly affected by various situation-specific sub-factors like velocity and position, but as discussed below, our findings are robust to the noise parameter, and so we do not explore these differences as they do not matter to our main point.

We note that the estimate of 31 pixels is an upper bound on the perceptual uncertainty in our main task, because in the actual main task the balls remained visible throughout, whereas in the noise estimation task participants did not see the balls when they responded. Importantly, greater perceptual uncertainty biases the results in favor of the Parallel Model, and so to the degree that this upper-limit perceptual noise is different from people's actual (smaller) perceptual noise in the main task, this would serve only to strengthen the support in favor of a Serial model. So, the Serial Model is much preferred to a Parallel Model under the upper-bound noise estimation used in the main text, and it is even *more* preferred under the lower uncertainty which likely existed in the main task (and see also the noise analyses below).

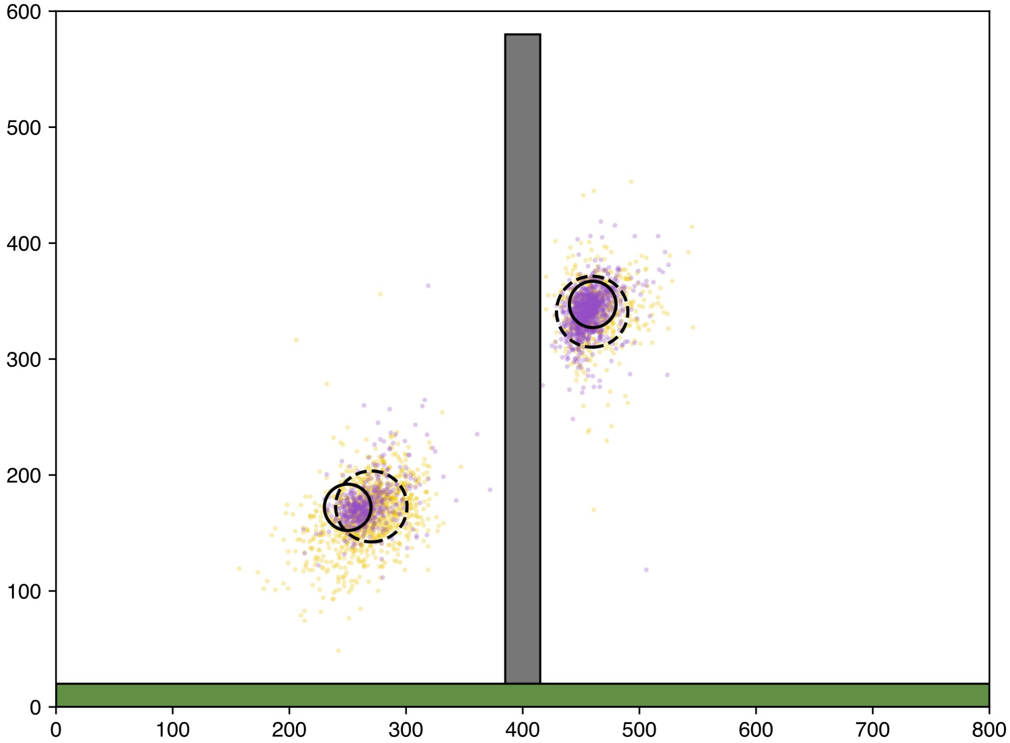


Figure S2: Noise estimation experiment (S1). A schematic view of the scene (axes show pixels), with all of the trials collapsed, and different scenes aligned relative to one example scene. Each dot is one response (aligned relative to the example scene). The first responses (within a given trial) are shown in purple, and the second responses (within a given trial) are in yellow. The solid black circles show the true final locations of the balls (with the size being equal to the balls' size), and the dashed circles show the SD of participants' responses (31 pixels), centered around the average response location for each ball.

Experiment S2: Two Objects in Minimally-Physical Scenarios

Our main experiments included a physical simulation task, where the to-be-imagined objects moved under gravity and could undergo collisions with a separating barrier. One might worry that the serial pattern we found was not a characteristic of mental simulation generally, but instead originates specifically from the physical reasoning involved in predicting the trajectories of the given task, and as opposed to the simplified stimuli used in many perceptual tracking tasks. As argued in the main text, we view a physically-realistic task as much more likely to allow for efficient processing, given the rich experience people have in everyday life with objects acting under gravity and colliding. However, to examine the robustness and generalization of our findings, we ran another experiment that involved mentally simulating objects that are moving along straight lines, in constant speed, with no collisions or gravity – as is commonly used in Multiple Object Tracking.

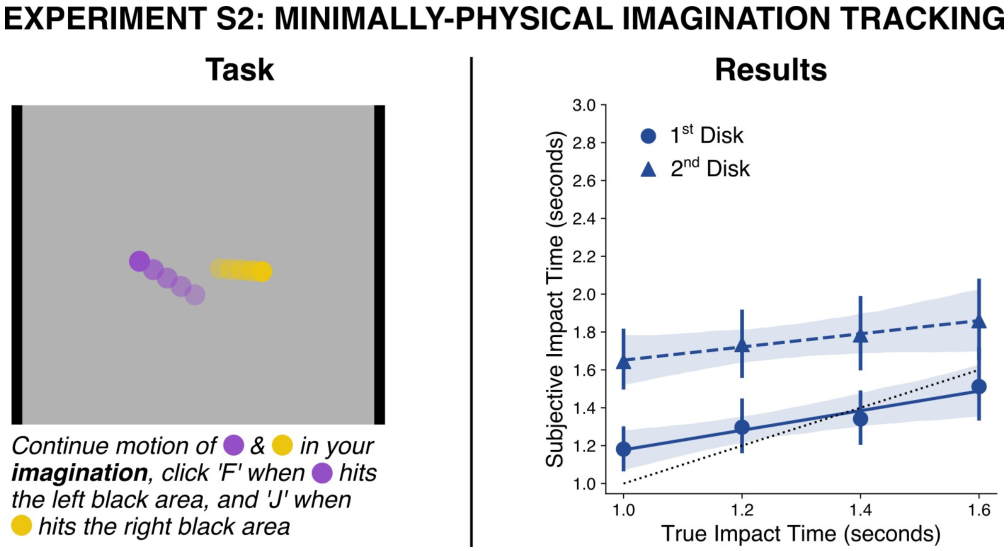


Figure S3: Task and results of Experiment S2, tracking two objects in imagination, with minimally-physical trajectories. Circles and triangles indicate mean responses for different true impact time and response order ('1st ball' refers to the ball participants responded to first, and '2nd ball' to the ball they responded to second), error bars show SEM, solid line shows best linear fit, shaded area is 95% CI, and dotted line shows hypothetical perfect performance (where the subjective impact time equals the true impact time), as reference.

Method

The task was similar overall to Experiment 1b: scenes showed two objects pausing mid-motion, and the task was to continue trajectories in the mind's eye and indicate when each object collides with a specific area. The difference was that the animations and description given to participants were altered in the following ways (see Figure S3, left; for a demo, see `https://jatos.mindprobe.eu/publix/GX5Rdu8q3VP`): We created videos using the same physics engine as before, but with gravity turned off. The scene background was gray, with two black rectangles spanning the height of the video placed on the right and left edges. Two colored disks were presented roughly in the center of the screen, and moved in straight lines and with a constant speed towards the right and left sides of the scene. Participants were asked to press the left side key when the left side disk collides with the left side wall, and the right side key for the right side disk and right side wall.

As in the main experiments (and following the pre-registered protocol), we recruited participants until we had 36 participants that passed the quiz and had a sufficient proportion of unique responses (see the main text), which required N = 69. We analyzed the results in the same way as in the main experiments.

Results

As can be seen in Figure S3 (right), tracking two objects in imagination using minimally-physical stimuli that are similar to conventional perceptual tracking tasks replicate the serial pattern of the main studies. Participants' response were linearly modulated by the true impact time, F(1.41, 49.40) = 26.44, $p < 0.001$, partial $\eta^2 = 0.43$; linear trend: t(105) = 8.79, $p < 0.001$. However, as in the more physical experiments, the second response happened much later than the first, F(1, 35) = 132.56, $p < 0.001$, partial $\eta^2 = 0.79$. Comparing the results to Experiment 1b, we found a significant interaction of Experiment with Response Order, F(1, 70) = 8.92, p = 0.004, partial $\eta^2 = 0.11$, driven by a larger effect in Experiment 1b. In Experiment S2, the average delay was 423 ms (CI for the intercept of the first response: [442, 880] ms, second response: [1,075, 1,532] ms), and the interaction between response order and the true impact time was not significant, F(2.37, 83.15) = 2.07, p = 0.12). In terms of fits, the Serial model explained 98% of the variance in responses, MSE = 0.001, while the Parallel model explained 27% of the variance, MSE = 0.04.

These findings demonstrate that our results are not due to the specific physical requirements imposed by the main task. Mentally simulating the movement of items appears to happen on a single-item basis in both complex and simple stimuli.

Experiment S3: Imagination Tracking with Increased Motivation

In our main tasks, we were interested in people's natural behavior without pushing them toward a specific tactic. As a result, it is in principle possible that participants adopted a serial simulation mode as a strategic choice, and that parallel simulation is possible when people 'try harder'. The main text presents several reasons to doubt the idea that our participants chose serial simulation to save on resources (and see also the individual performance analysis that shows the serial pattern is not due to only a few participants, instead being an extremely robust finding at the individual level). Here, we describe an additional experiment where we tried motivating participants to maximize their performance for monetary reward, to test whether this incentive can push their responses towards parallel simulation.

Method

The task and stimuli were identical to Experiment 1b, with the only difference being a monetary bonus (see Figure S4, left; for a demo, see `https://jatos.mindprobe.eu/publix/OMaFIruJAbS`). During the instructions phase, we told participants that the closer their responses are to the true impact times, the higher they will score towards an additional monetary bonus to a maximum of $1 (with the baseline, non-bonus payment being $1.6). We reminded participants of the bonus throughout the experiment, asking them to perform the task as precisely as they can in order to win the highest bonus. Because we only care about potential changes in overall performance as a function of motivation, after finishing the experiment, all of the participants were informed that they will receive a bonus of $.85. To prevent participants learning various heuristics, we did not provide feedback throughout the task.

As in the main experiments (and following the pre-registered protocol), we recruited participants until we had 36 participants that passed the quiz and had a sufficient proportion of unique responses (see the main text), which required N = 55. We analyzed the results in the same way as in the main experiments.

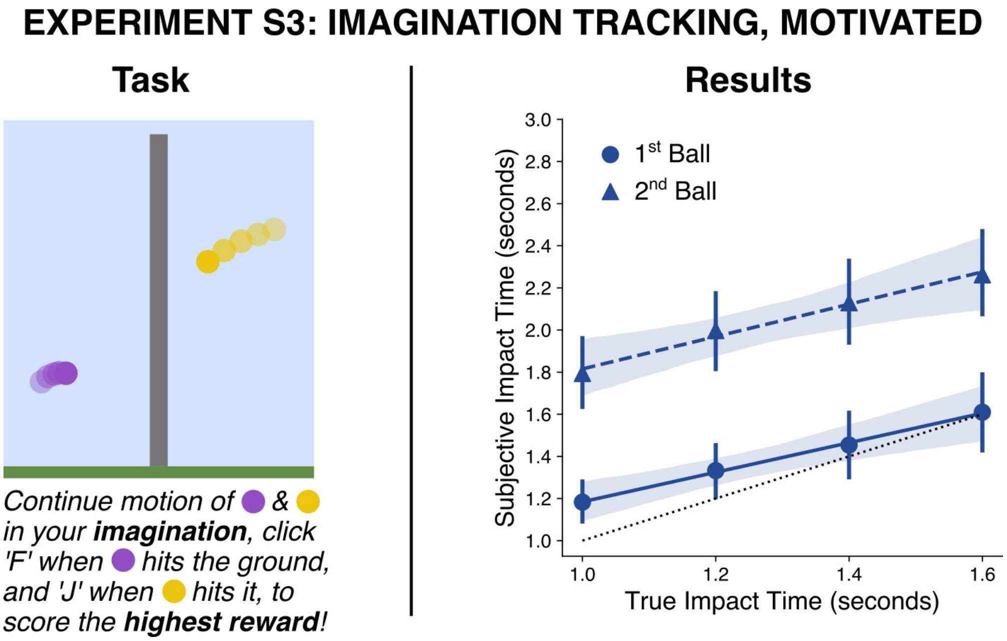


Figure S4: Task and results of Experiment S3, tracking two objects in imagination, with monetary bonus. Circles and triangles indicate mean responses for different true impact time and response order ('1st ball' refers to the ball participants responded to first, and '2nd ball' to the ball they responded to second), error bars show SEM, solid line shows best linear fit, shaded area is 95% CI, and dotted line shows hypothetical perfect performance (where the subjective impact time equals the true impact time), as reference.

Results

Motivating participants to perform as accurately as possible did not change the pattern of responses. As shown in Figure S4, right, we closely replicated the results of Experiment 1b. Participants' response were linearly modulated by the true impact time, F(1.36, 47.64) = 65.25, $p < 0.001$, partial $\eta^2 = 0.65$; linear trend: t(105) = 13.96, $p < 0.001$. However, as in Experiment 1b, the second response happened much later than the first, F(1, 35) = 235.39, $p < 0.001$, partial $\eta^2 = 0.87$. Comparing the results to Experiment 1b, we found no interaction of Experiment with the Response Order effect, F ¡ 1, suggesting highly similar patterns of results. In Experiment S3, the average delay was 651 ms (CI for the intercept of the first response: [285, 682] ms, second response: [797, 1,298] ms), and the interaction between response order and the true impact time was not significant, F ¡ 1. In terms of fits, the Serial model explained 99% of the variance in responses, MSE = 0.0007, while the Parallel model explained 30% of the variance, MSE = 0.09.

These results suggest the serial capacity limit is strict, in that people cannot simply adjust their imagination to simulate independent objects in parallel through effort and motivation.

section*Experiment S4: Imagination Tracking with Occlusion

In our main tasks, the items freeze mid-motion and participants are asked to complete their movement in the imagination. This obviously deviates from how scenes naturally unfold in the real world. Could it be that the serial pattern we observed does not reflect a capacity limit on mental simulation per se, but is the result of some disruption to everyday

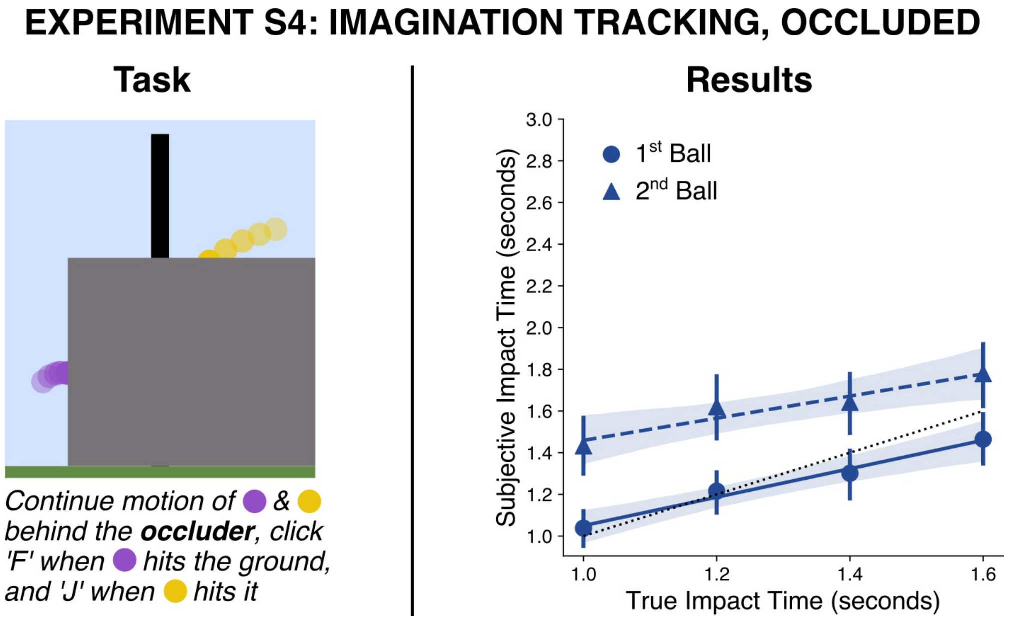


Figure S5: Task and results of Experiment S4, tracking two objects in imagination, with occlusion. Circles and triangles indicate mean responses for different true impact time and response order ('1st ball' refers to the ball participants responded to first, and '2nd ball' to the ball they responded to second), error bars show SEM, solid line shows best linear fit, shaded area is 95% CI, and dotted line shows hypothetical perfect performance (where the subjective impact time equals the true impact time), as reference.

spatiotemporal tracking (see 2)? To test this, we conducted an additional experiment where the items gradually disappeared behind an occluder, instead of freezing.

Method

The task and stimuli were identical to Experiment 1b, except as stated below. First, the balls did not freeze but continued to move. Second, a gray rectangle occluded much of the bottom part of the scene (Figure S5, left; for a demo, see `https://jatos.mindprobe.eu/publix/Brm7G1m7ogT`). The dimensions of the occluder on each trial were chosen so that the balls disappeared behind it after 500 ms of movement. It spanned vertically from the top of the ground to the bottom point the ball that collides with the wall reached at 500 ms, and horizontally from the side of the video on the colliding ball's side (to hide how it rolls on the ground post-hit) to the innermost point the hyperbole ball reached at 500 ms. Third, the wall was presented in black instead of gray so it is salient against the gray 'screen'. Fourth, the instructions, example trials, and quiz questions were updated to explain the occlusion. Participants first saw unoccluded trials during practice, and then occluded trials.

As in the main experiments (and following the pre-registered protocol), we recruited participants until we had 36 participants that passed the quiz and had a sufficient proportion of unique responses (see the main text), which required N = 58. We analyzed the results in the same way as in the main experiments.

Results

As can bee seen in Figure S5, right, mentally simulating items that disappear in a natural way behind an occluder still produced a serial pattern, similarly to the results of Experiment 1b where the items froze mid-motion. Participants' responses were linearly modulated by the true impact time, F(2.02, 70.85) = 158.91, $p < 0.001$, partial $\eta^2 = 0.82$; linear trend: t(105) = 21.43, $p < 0.001$. However, as in the experiments that involved items freezing, the second response happened much later than the first, F(1, 35) = 141.28, $p < 0.001$, partial $\eta^2 = 0.8$. Comparing the results to Experiment 1b, we found a significant interaction of Experiment with Response Order, F(1, 70) = 15.65, $p < 0.001$, partial $\eta^2 = 0.18$, driven by a larger effect in Experiment 1b. In Experiment S4, the average delay was 364 ms (CI for the intercept of the first response: [226, 509] ms, second response: [759, 1,097] ms), and the interaction between response order and the true impact time was significant, F(2.36, 82.64) = 4.78, p = 0.007. Examining model fits showed that the Serial model explained 96% of the variance in responses, MSE = 0.002, while the Parallel model explained 46% of the variance, MSE = 0.03.

These results suggest the serial capacity limit does not simply reflect the unnatural freezing we used in the main experiments, and can be observed also for items that are simply occluded. Occlusion did mitigate the serial effect somewhat, and this might reflect any of a number of factors, such as the greater spatial and temporal predictability regarding the switch from perception to imagination, or the greater familiarity of the more ecological scenarios. Indeed, this might point to interesting strategies people employ in the real world to handle a complex world with such a limited capacity to mentally simulate events. Given how common occlusion is in the real world, the fact we could replicate the Serial pattern in Experiment S4 suggests that the single object bottleneck arises under naturalistic simulation conditions as well, corroborating the importance of the present findings.

Testing the Effect of Noise in the Models

The models in the main text use a perceptual noise parameter that we independently estimated in the task described in the previous section. As mentioned, the perceptual noise estimated using this task is an upper limit on the expected noise in the main task. In addition to this, we also carried out a systematic exploration of perceptual noise levels. We created variants of our Serial and Parallel models, by varying a noise parameter D, equivalent to the SD of a symmetric two-dimensional Gaussian distribution centered around each ball's location at the beginning of the simulation, and measured in units corresponding to the radius of the balls. Our analysis varied this parameter from $D = 0$ (no perceptual noise), through $D = 1.5$ (the upper bound found through human participants), and up to $D = 10$ (ten times the radius of the balls, or 200 pixels).

As the noise increases, the fit of the Parallel Model gradually improves, because the noise randomly makes one ball reach the ground later and produces a delay between the responses. Still, the Serial Model outperforms the Parallel Model for nearly all parameter values considered, including *vastly* outperforming it in the relevant ranges as estimated by participant uncertainty (see the Table). The only parameter setting in which the Parallel

Table 1: Model fits in terms of explained variance (R2) and MSE for different noise levels, and percentage of invalid location trials (i.e., where the noise moved a ball across the wall or placed it outside the scene borders.). For the model fits, the winner model is in bold.

	% Invalid Locations		**Parallel Model Fit**		**Serial Model Fit**	
Noise SD (D)	**Parallel**	**Serial**	**R**	**MSE**	**R**	**MSE**
0	0.00	0.00	0.12	0.103	**0.97**	**0.004**
0.5	0.00	0.00	0.13	0.101	**0.96**	**0.004**
1	0.01	0.02	0.15	0.100	**0.96**	**0.004**
1.5	0.03	0.08	0.21	0.093	**0.97**	**0.004**
2	0.12	0.13	0.31	0.080	**0.96**	**0.004**
2.5	0.20	0.21	0.31	0.081	**0.96**	**0.005**
3	0.26	0.26	0.44	0.066	**0.97**	**0.004**
3.5	0.30	0.33	0.51	0.057	**0.97**	**0.004**
4	0.34	0.33	0.54	0.053	**0.95**	**0.006**
4.5	0.39	0.41	0.68	0.037	**0.95**	**0.006**
5	0.49	0.45	0.71	0.033	**0.95**	**0.006**
5.5	0.52	0.46	0.77	0.027	**0.97**	**0.004**
6	0.52	0.49	0.79	0.025	**0.98**	**0.003**
6.5	0.52	0.56	0.84	0.018	**0.96**	**0.004**
7	0.64	0.57	0.91	0.010	**0.95**	**0.005**
7.5	0.62	0.60	0.75	0.030	**0.92**	**0.009**
8	0.70	0.62	0.94	0.007	**0.94**	**0.007**
8.5	0.71	0.72	0.90	0.011	**0.95**	**0.006**
9	0.68	0.72	**0.97**	**0.004**	0.94	0.007
9.5	0.79	0.74	0.84	0.019	**0.97**	**0.003**
10	0.78	0.76	0.82	0.022	**0.90**	**0.012**

Model outperforms the Serial model ($D = 9$, $R^2 = 0.97$ vs. $R^2 = 0.94$) is at a noise level roughly 6 times larger than the upper bound we found in a task where the balls disappear, is absurdly large (a ball of this size would take up a third of the scene), and produces absurd location values: over two thirds of the noisy locations for the balls either cross the wall in the scene, or are beyond the scene's borders altogether.

We also conducted an analysis that specifically targets the influence of noise on model predictions in a more fine-grained manner. We calculated the unique contribution of the noise in the simulation by subtracting the true difference in the true impact times Δ_{gt} from the actual found response time difference Δ_h. In the absence of perceptual noise, the Parallel Model predicts values of 0 regardless of the true difference in impact time (all of the delay comes from the actual difference Δ_{gt} between the balls). By contrast, the Serial Model predicts a constant delay of 800 ms between the responses (because the second simulation starts anew once the first ball hits, and 800 was the average duration of the remaining movement across the different scenarios). So, with larger differences between impact times a larger value is subtracted and the trend declines linearly.

The results of applying this analysis to the noisy variants of the Parallel Model appear in

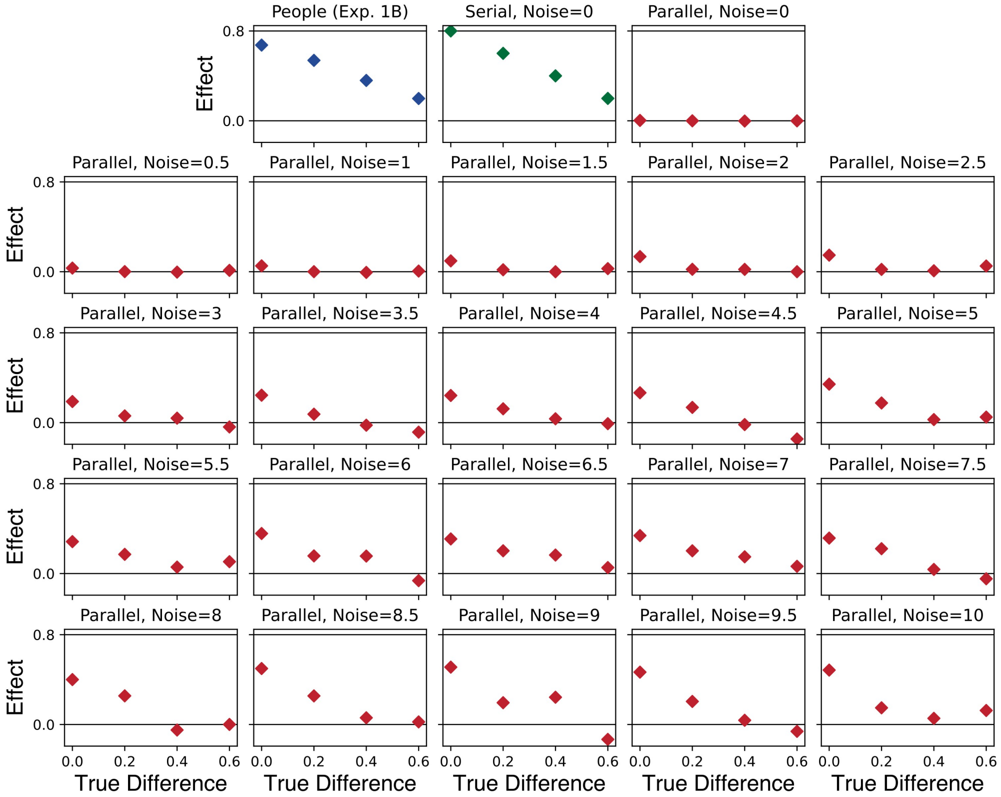


Figure S6: The effect of noise on the Parallel Model predictions. The top row shows, for comparison, participants' results in Experiment 1b, and the predictions of the noiseless Serial and Parallel Models. As the noise increases, the Parallel model predicts a larger delay between the balls, but this effect is more pronounced when the true difference between them is small (when all of the difference in impact time is attributed to noise) then when it is large.

Figure S6: As the noise grows, the Parallel Model begins to predict a decreasing trend, but it is too small while the balls have a large difference between them, to the point that some of the effect becomes negative. Participant behavior in Experiment 1b closely followed the Serial Model's predictions not only for the overall pattern, but also for this specific effect.

In summary, the more fine-grained noise-analysis shows that the main finding (Serial simulation is a better explanation of people's behavior than Parallel simulation) holds for a wide range of noise values, and matches the qualitative trends as well.

Interleaved Serial Models

The Serial Model we focused on in the main text unfolds the complete trajectory of one object before it starts unfolding the trajectory of the second object. But there is a more nuanced possibility: perhaps people interleave the mental simulation of two objects, such that first an object is moved forward for S steps, then the simulation switches to the other object for an additional S steps, then the simulation cycles back to the first object, and so

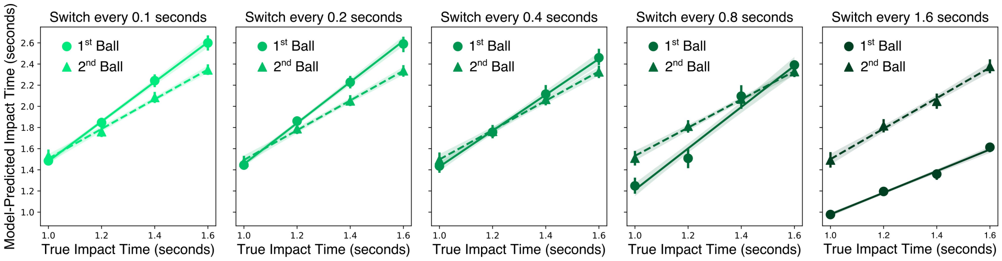


Figure S7: Model predictions in Experiment 1b (tracking two balls in imagination), for different values of the switching parameter. Circles and triangles indicate mean responses for different true impact time and response order, error bars show SEM, colored lines show best linear fit, and shaded area is 95% CI.

on. At first glance, such a model might seem to lie in the middle of the Serial and Parallel Models, such that a value of $S = 1$ is equivalent to the Parallel Model, and as S grows we approach the Serial Model of the main text. As the analysis below shows, this is not the case.

To test the influence of interleaving, we created 5 additional models, each with a different S parameter: switching every 100, 200, 400, 800 , or 1600 ms of simulation. Setting $S = 1600$ is roughly equivalent to the Serial Model considered in the main text.

The results, shown in Figure S7, reveal a pattern that might be unintuitive at first. As the interleaving window becomes smaller, the first and second responses become more similar to each other. But, this does not reflect a faster second response (as in the Parallel Model) but a slower first response. This is because the first ball only manages to hit the ground after several switches. Such a prediction does not line up with our data, in which we found that the first response in Experiment 1b (2-balls imagination tracking) is highly similar to the first-and-only response in Experiment 1a (1-ball imagination tracking).

Beyond this, with small interleaving windows (smaller S), we find that larger values of the true impact times can lead to average second responses that are *faster* than the average of first responses. This seemingly paradoxical effect simply reflects an uneven distribution of responses across the true impact times. To see this, consider a ball B_1 with the longest true impact time of 1600 ms, and a short interleaving window (small S): if the second ball B_2 has a shorter true impact time than B_1, it will almost certainly reach the ground first. At that point, the simulation is 'released' to move only B_1. Such a situation includes most of the scenarios. As the difference between the true impact times of the two balls becomes smaller, the moment when the ball B_1 'breaks off' happens later, producing a crossing point in the figures. This 'crossing effect' is important for two reasons: First, even in the limit of minimal S, this effect is not produced by the Parallel Model, showing that an interleaved model with maximal switching is *not* the same as the Parallel Model. Second, this effect is not found in participant data, suggesting again that the Serial Model considered in the paper (with maximal S) more closely matches human behavior.

We believe that an interleaved Serial Model may be relevant for situations beyond the main task considered in the paper, in which two or more objects may interact with one another (e.g. in the case of collisions). However, given that the interleaved model makes the two qualitative predictions detailed above, two predictions that do not line up with the full

Serial Model of the main text, we find there is no evidence for the interleaved model in our data.

Supplemental Model Fitting

In the main text, we present the results of one type of model comparison for the Serial vs. Parallel Tracking Models: linear fits between Model-predicted impact times, and the average subjective impact time given by participants. Here, we detail two additional analyses. To summarize up front, the results of both of these separate approaches agree: Participant behavior in Experiment 1b (two separate objects in imagination) is better explained by the full Serial model, while participant behavior in Experiment 3 (two objects in imagination with strong grouping) lies somewhere between the the Serial and Parallel models.

Noisy simulation, individual data fit. This analysis fits the noisy simulation models to the individual data, aggregated across all participants, and still fitting overall a and b parameters of a linear fit. The results in Experiment 1b were still in line with serial simulation in imagination: the Serial Model explained 24% of the variance in responses, MSE = 0.36, while the Parallel Model explained 5% of the variance, MSE = 0.45. Adding grouping cues (Experiment 3) pushed the results towards parallel simulation: the Serial Model explained 10% of the variance in responses, MSE = 0.81, and the Parallel Model explained 12% of the variance, MSE = 0.79.

Noiseless model fits. As noted in the fine-grained noise-analysis above, the Serial Model is preferred to the Parallel Model in condition of $D = 0$ as well. To detail this a bit further, for Experiment 1b (two objects in imagination) the Serial Model explained 97% of the variance in responses, MSE = 0.004, and the Parallel Model explained 11% of the variance, MSE = 0.1. Thus, the Serial Model does not need any level of noise to account for participants' performance. In addition to that analysis, we considered noiseless models for Experiment 3 (two objects in imagination with grouping cues), and found that the grouping cues pushed the results more towards parallel simulation: the Serial Model now explained 69% of the variance in responses, MSE = 0.04, and the Parallel Model explained 78% of the variance, MSE = 0.028.

Response Order Accuracy

Our Serial Model assumed that the choice of the first object to simulate is arbitrary and selects it at random. However, if serial simulation is an accurate description of human imagination, it is likely that people's object selection is biased in some way. While this was not the main focus of our studies, we considered this possibility by analyzing the frequency at which participants responded in the correct order, for scenarios in which the balls differed in their true impact time. In a parallel simulation, the chance of responding in the correct order should be close to 100%. In a serial simulation, if people respond in the correct order 100% of the time, it means they know ahead of time which ball would hit the ground first

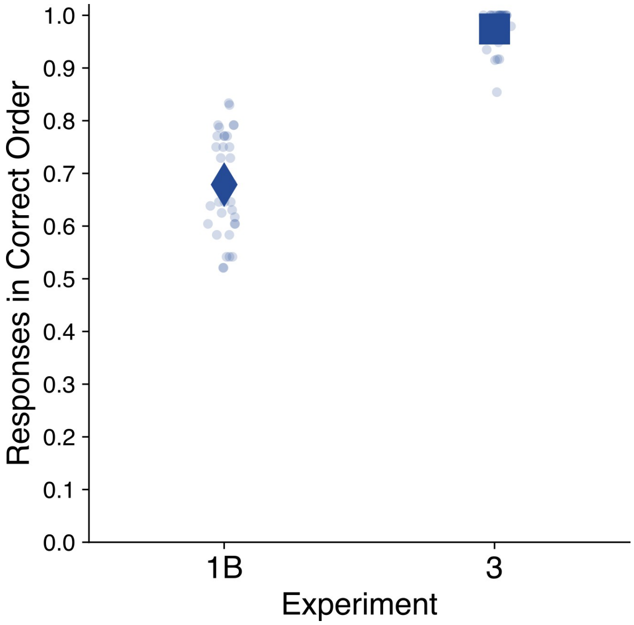


Figure S8: Percentage of trials where participant responded in the correct order, by experiment: raw individual (small markers) and average (large markers) results.

(presumably through a process other than simulation). If people respond in the correct order 50% of the times, it would suggest that they pick the first ball to simulate at random. An intermediate value would suggest that people have some imperfect heuristic for picking the first ball to simulate (e.g., based on height or velocity). This last option would also be in line with recent work on the combination of heuristics and simulation (e.g., 3).

So, we can use the accuracy of *order* of collision as a metric for whether people's choice of first object is biased. Figure S8 shows the results of such an analysis of order accuracy, for the two experiments that included two objects in the imagination. In Experiment 1b, the average order accuracy was 68% (SEM=9%), suggesting that people neither choose the first ball perfectly nor at random, but likely use some imperfect heuristic. In Experiment 3, the average order accuracy was 97% (SEM=3%), suggesting that people can use the height of the balls as a useful heuristic, which indeed in this situation fully determined the order in which the balls landed on the ground.

Individual Performance with a Single Object

In Experiment 1a, when people tracked the trajectory of a single object in imagination, we found that responses were linearly modulated by the true impact time, suggesting a high sensitivity to ground truth physics. Here, we show that this finding is not an artifact of averaging: it is found in the individual responses of almost all participants, as Figure S9 shows. This demonstrates that people are tuned to even subtle adjustments of physical properties, as the differences between different impact times were as small as 200 ms.

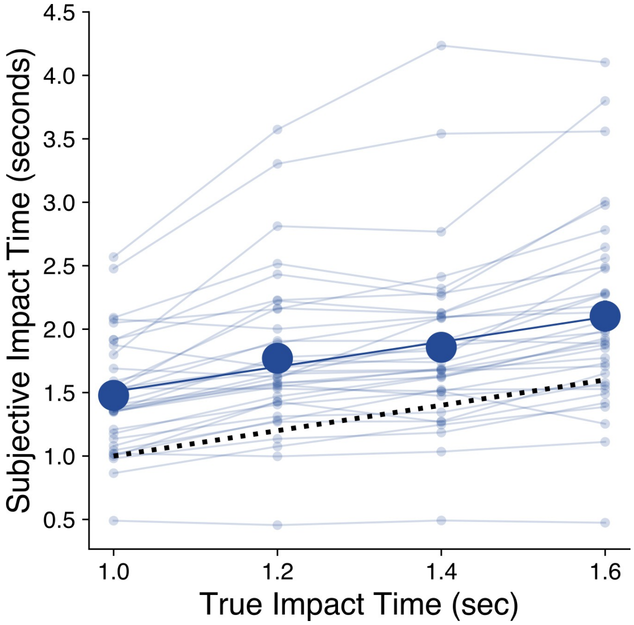


Figure S9: Individual performance in Experiment 1a, with a single object in imagination, by the True Impact Time: raw individual (small markers) and average (large markers) results. The dotted black line shows hypothetical perfect performance (where the subjective impact time equals the true impact time), as reference.

References

[1] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum, "Simulation as an engine of physical scene understanding," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 18327–18332, Nov. 2013.

[2] B. J. Scholl and Z. W. Pylyshyn, "Tracking Multiple Items Through Occlusion: Clues to Visual Objecthood," *Cognitive Psychology*, vol. 38, pp. 259–290, Mar. 1999.

[3] K. Smith, P. Battaglia, and J. Tenenbaum, "Integrating heuristic and simulation-based reasoning in intuitive physics," preprint, PsyArXiv, July 2023.