"Just In Time" Representations for Mental Simulation in Intuitive Physics

Tony Chen, Kelsey Allen, Samuel Cheyette, Joshua B Tenenbaum, Kevin A Smith
{thc, krallen, cheyette, jbt, k2smith}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract

Many models of intuitive physical reasoning posit some kind of mental simulation mechanism, yet everyday environments frequently contain far more objects than people could plausibly represent with their limited cognitive capacity. What determines which objects are actually included in our representations? We asked participants to predict how a ball will bounce through a complex field of obstacles, and probed working memory for objects in the scene that were more and less likely to be relevant to the ball's trajectory. We evaluate different accounts of relevance and find that successful object memory is best predicted by how frequently a ball's trajectory is expected to contact that object under a probabilistic simulation model. This suggests that people construct representations for mental simulation efficiently and dynamically, on the fly, by adding objects "just in time": only when they are expected to become relevant for the next stage of simulation.

Keywords: physics; representations; resource rationality

Introduction

Imagine throwing a Frisbee to a friend while you are both standing near a grove of trees. Figuring out where your throw will go clearly requires representing the Frisbee and your friend. But you might also need to account for the nearby trees, as the Frisbee could hit an overhanging branch, or one of the trunks. You might not need to consider the trees further into the grove, as the Frisbee is unlikely to make it that far. This is emblematic of a challenge that arises any time we must predict or act in the world: there are always a large number of objects around us, and we do not have the cognitive resources to model all possible interactions and effects (Ludwin-Peery, Bramley, Davis, & Gureckis, 2021; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). Instead, we must decompose our environment into a representational form that is useful for prediction, simulation, and action; a process that often involves picking out a subset of perceptual input to attend to and ignoring the rest. But how does the mind determine what to represent and what can be thrown away? Here we propose that this involves adding objects to our representations "just in time" as they are needed for our simulations, thereby ignoring objects that do not become relevant.

A large body of work in attention and visual working memory has shown that people are efficient with their allocation of cognitive resources, distributing attention to features and objects in the environment that are relevant for the task at hand (Bates, Lerch, Sims, & Jacobs, 2019; Bates & Jacobs, 2020; Emrich, Lockhart, & Al-Aidroos, 2017). However, these notions of relevancy primarily take the form of perceptual statistics that can be easily learned through associative learning mechanisms. But there are many situations where relevancy cannot be estimated based on the statistics of past experiences, and instead is dependent on the outcome of a process that uses the mental representations in question. Do we still flexibly allocate cognitive resources in these cases, and if so, how do we do so?

The domain of intuitive physics is well suited for studying this question. The world is a complex place, where precisely determining what happens in the future requires interactions between many more objects than could reasonably held in working memory. Yet recent research suggests that naturalistic physical predictions are based on relatively accurate simulations (Ahuja & Sheinberg, 2019; Smith, Battaglia, & Vul, 2018), even in scenarios where there are a large set of relevant objects, like predicting whether and how a stack of blocks will fall (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Zhou, Smith, Tenenbaum, & Gerstenberg, in press). The complexity of these scenarios has led to critiques of simulation as requiring infeasibly detailed representations (Ludwin-Peery et al., 2021). But others have begun to explore various approximations the mind might make in its representations in order to efficiently make physical predictions (Ullman et al., 2017), such as representing objects with coarse shape approximations (Li et al., 2022), or even dropping objects from our representations entirely (Bass, Smith, Bonawitz, & Ullman, 2022). But even with a catalogue of possible simplifications, the question of *how* the mind chooses the right approximation is an open question (Davis & Marcus, 2015).

In recent years, the framework of resource rationality (Lieder & Griffiths, 2020) has been proposed to explain trade-offs between accuracy and efficiency in decision making (Bhui, Lai, & Gershman, 2021), planning (Callaway et al., 2018), and cognitive control (Musslick, Saxe, Hoskin, Reichman, & Cohen, 2020). More recently, Ho et al. (2022) extended this framework to explain trade-offs in the complexity of our *representations*. They showed that people construct reduced representations of obstacles and objects in a grid world navigation task, and that the objects are more likely to be included in this reduced representation if they would be relevant to the navigation task.

Here we consider what information people use to define *relevancy*, and the implication this has for how reduced representations are constructed. Ho et al. (2022) focuses on a value-based notion of relevancy – the probability that a given obstacle is included in a representation depends on the difference of reward obtained when planning with and without that object included in the representation (Fig. 1, left). This presents a challenge for algorithmic implementations of the construal process, since a plan is required to determine which representation to use for planning. This is a general chal-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

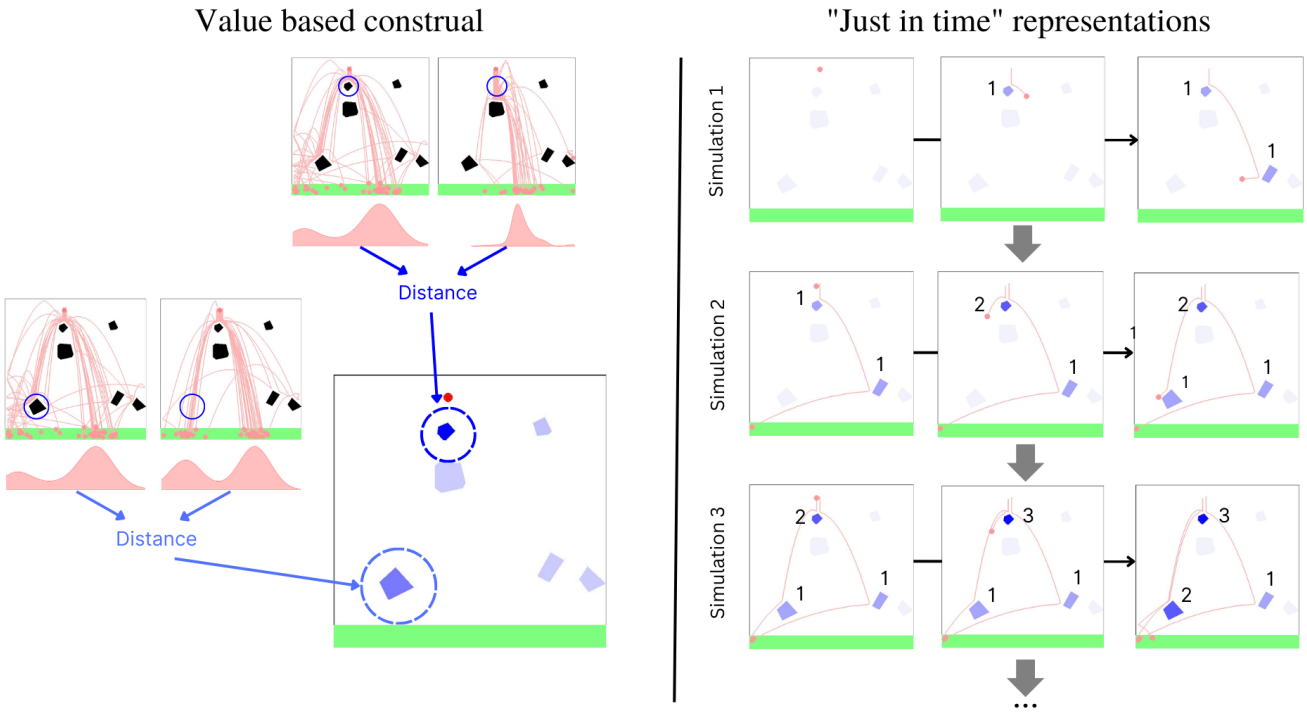

Figure 1: **Left:** The *Value-Based Construal* model. The relevance of each object is determined as the difference that object would make to the outcome, calculated as the divergence between the simulated outcome and the outcome if that object had been different or removed. Objects with larger outcome divergence (such as the top object) are more likely to be included in the representation than objects with less outcome divergence (such as the object in the lower left). **Right:** The *"Just in Time" Representations* model. The representation is built up iteratively, starting with a sparse representation and simulating until an object becomes relevant to the simulation, then adding that representation "just in time" for it to be used in simulation. In this example, the numbers index representation strength for each object: during Simulation 1 the ball collides with the top and bottom-right objects, and so these objects will be added to the representation to resolve these collisions. In Simulation 2, the ball collides with the top object, strengthening its representation, then with the bottom left-object, necessitating the representation of that object. In Simulation 3, the ball again collides with the top and bottom-left objects, further strengthening both those representations.

lenge faced in resource rational systems (Gershman, Horvitz, & Tenenbaum, 2015) that often suggested to be solved by offline learning of efficient strategies (Erev & Barron, 2005; Lieder & Griffiths, 2017; Siegler, 1999).

But we also consider an alternate "just in time" model of constructing representations. Under this account, objects are only added to one's representation when they are needed for simulation: people start with an extremely impoverished representation and simulate until an object they see in the environment would affect the outcome of the simulation (e.g., because another object is about to collide with it), at which point they include that object into their representation. The process of representing objects as needed continues iteratively, with representations growing stronger for objects that are relevant to simulation more often (Fig. 1, right). This model would also explain why people look more often at the locations in scenes where they expect collisions to occur (Beller, Xu, Linderman, & Gerstenberg, 2022; Crespi, Robino, Silva, & de'Sperati, 2012): they are attending selectively to objects that they are including in their representations. Under this account, relevance is not a function of how much the inclusion of an object will affect the final prediction, but rather just how likely that object is to be impact simulation at all.

We take a first step towards adjudicating between these two theories of representational construction for physical reasoning. Because we have well described, probabilistic models of physical simulation (Battaglia et al., 2013; Smith & Vul, 2013), we study how representations change with graded differences in various forms of relevancy. To investigate the principles that govern human reduced representations in physical tasks, we test participants in a physical simulation paradigm, in which participants were asked to predict the position of a ball as it falls down a board filled with obstacles. On critical trials, after making their predictions, participants were presented with a task in which their memory of specific obstacles was tested. Crucially, the probed obstacles varied in how relevant they were to the simulation of the ball, ranging from not colliding at all with the ball to colliding with the ball every time.

We find that peoples' representations are tied to the outcomes of simulations in these environments, with objects that are predicted to almost certainly be in the path of the ball being remembered best, and people having the poorest memory for objects away from the ball's path. We additionally find that memory for these objects is not as well explained by alternate measures of relevance that measure how transformations or deletions of the object would affect the expected end state of the ball. Our results provide evidence that people

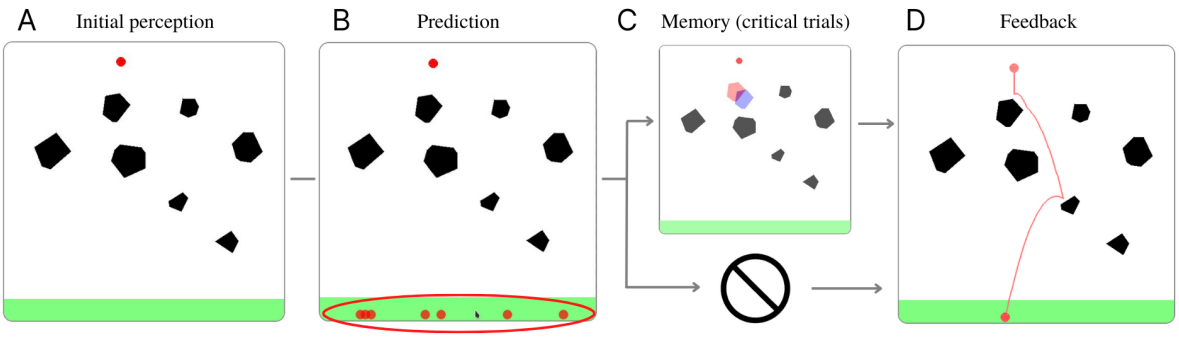

Figure 2: Schematic of the experimental paradigm. **(A)** Participants first perceive a scene, then **(B)** indicate ten locations where they expect the ball to land. **(C)** After this, in one third of the trials, participants are given a memory test: the probed object is shown alongside a shifted version of the same object, and participants must decide which object was in the original position from the previous scene. **(D)** Finally, at the end of each trial a video of the actual trajectory of the ball is shown.

use mental simulation over a compressed representation of the world that omits irrelevant objects. We hypothesize that in domains such as planning and physics, these representations are constructed by an interleaved process of forward simulation and "just in time" construal, gradually building representations and predictions to efficiently model the world.

Experiment 1

In order to test whether people's object representations are constructed depending on how that object contributes to a simulation, we test participants in a Plinko style prediction paradigm (Gerstenberg, Siegel, & Tenenbaum, 2021; Beller et al., 2022; Fig. 2) where they are asked to predict the trajectory of a ball as it falls down through a set of obstacles. However, one third of the trials are critical memory trials, in which participants are shown a scene with the original object and an additional distractor object constructed by applying a shift to the original object, and asked to judge which object was the one that they saw in the original scene. We constructed the memory task so that the shifted object has varying degrees of relevance to the outcome of the simulation – from objects that are always relevant because they are positioned directly in the path of the ball, to never-relevant objects that are far off to the side, and in between. Here we test whether peoples' memory for the correct position of objects is related to their relevance.

Participants. We recruited 220 participants from the Prolific research platform. The task took approximately 25 minutes, for which participants were compensated $6.25. We did not exclude any participants from our analyses.

Procedure. Participants initially viewed a still image of a ball suspended above a series of obstacles (Fig. 2A), and were asked to predict where the ball would touch the ground (Fig. 2B). We asked them to provide a range of predictions by clicking 10 times on the ground where they believed the ball would land; participants were instructed that they could click in the same location to indicate more confidence that the ball would land there.

On two-thirds of the trials (the *filler* trials), participants then would observe the actual trajectory that the ball took as it bounced down the Plinko system. At the end, participants earned points for the accuracy of their predictions. The grid of possible end states was discretized into 10 bins, and points were awarded proportionally to the bin distance between each prediction and the true outcome bin. These points were only used as a motivator and did not affect compensation.

The remaining third of the trials were critical memory trials. On these trials, immediately after indicating their predictions, the screen was masked and another Plinko scene was shown with the marked object colored in either red or blue, and an additional distractor object marked with the opposite color (Fig. 2C). The colors assigned to either the original object and the distractor were randomized across trials. Participants were asked to judge which object was actually present in the original scene, indicating their response on a slider ranging from "definitely sure the red object is in the correct position," to "definitely sure the blue object is in the correct position"; this response was translated to an integer from 0 to 100 and normed so that 100 indicated full confidence in the correct object, while 0 indicated full confidence in the distractor object. After the memory task, participants were shown the trajectory and the points earned, as was done in the filler trials.

Stimuli. Plinko boards were created by randomly generating between 4 and 10 polygon obstacles with between 4 and 7 sides, and placing them in random locations on the screen. The ball was then positioned at the top of the screen, and randomly placed horizontally in the center 20% of the world. We constructed 48 different trials: 32 filler trials and 16 critical memory trials. The filler trials were constructed only subject to the constraint that the ball did not get stuck on the obstacles. All participants judged all filler trials.

The critical memory trials were constructed to produce four qualitatively different object types, depending on how often the ball collided with that object in a set of noisy simulations (Fig. 3). The *Collision Early* objects were assumed to be most relevant, as they were objects that the ball collided with in all simulations, and were usually the first or second object that the ball touched. The *Collision Late* objects were

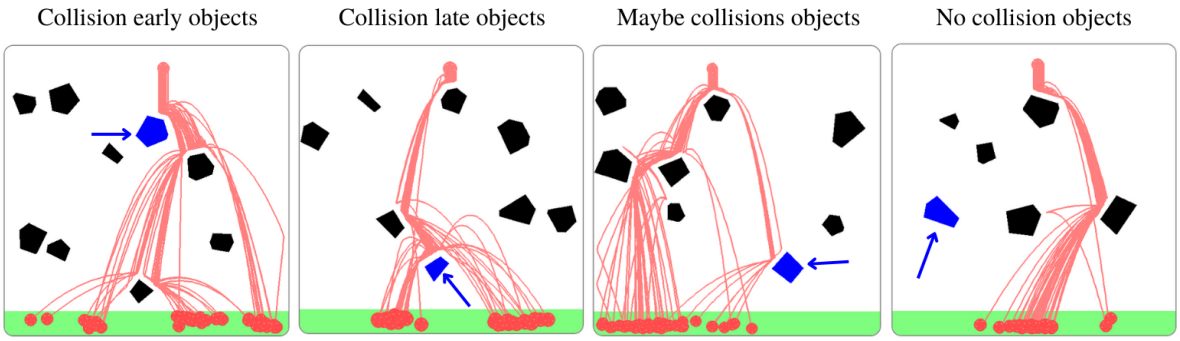

Figure 3: The four types of objects probed in the memory task. Red lines represent the paths traced by 40 different noisy simulations. The *collision early* objects are positioned underneath the release point of the ball, so that it will nearly certainly have an effect on the ball's trajectory. The *collision late* objects are lower in the scene, but are still in the path of the majority of simulations, so are still very likely to impact the trajectory. The *maybe collision* objects are approximately the same height as the collision late objects, but are only along the path of a portion of the simulation, making them less relevant to trajectory. The *no collision* objects are off to the side of all simulated paths and are expected to have no effect on the ball's trajectory.

assumed to be somewhat less relevant, as they were objects that over 95% of simulations touched but were late in the causal chain, with other collisions occurring first. The *Maybe Collision* objects were assumed to be even less relevant, occurring late in the causal chain but only touched by the ball in 40-60% of all simulated paths. Finally, the *No Collision* objects were assumed to be irrelevant for prediction, as the ball never touched these objects in simulation. The simulator and amounts of noise were identical to those used in Allen, Smith, and Tenenbaum (2020).

Critical memory trials were constructed to have objects of three different types: Collision Early, No Collision, and either Collision Late or Maybe Collision (since scenes with both of those blocks are very rare). We constructed 16 different critical trials – 8 with Collision Late, 8 with Maybe Collision – and provided them to participants in a counterbalanced way such that each participant made four judgments about each of the four collision types, observing each critical trial once.

Results Predictions of where the ball would end up were mainly used to calibrate simulation models; see Fig. 5 for examples.

Consistent with our expectations, we find that memory does vary across the different relevance types ($\chi^2(3) = 57$, $p = 2.4 * 10^{-12}$),[1] with the highest memory for Collision Early objects (71.5, 95% CI = [68.4, 74.6]), followed by Collision Late objects (63.3, 95% CI = [59.3, 67.3]), then Maybe Collision objects (56.3, 95% CI = [52.3, 60.3]), and finally No Collision objects (50.3, 95% CI = [47.2, 53.4]). Thus we find that the likelihood or detail of objects within participants' representations is related to gross differences in those objects' relevance to simulation.

Experiment 2

While the Experiment 1 results show that people's representations are sensitive to general differences in relevance, the discrete notion of relevance used precludes us from investigating a potentially graded effect of relevance at the object level. We therefore ran a second experiment to explicitly gather peoples' judgments of whether the ball will contact each target object, as a measure of relevance.

[1]For these analyses we use a random effects model with intercepts varying by subject, and by relevance type nested within trial.

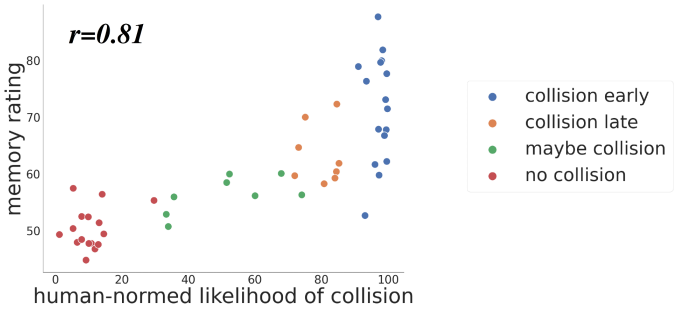

Figure 4: Comparison of human predictions of collision likelihood from Experiment 2 (x-axis) vs memory ratings from Experiment 1 (y-axis). Each point is a separate critical object; color indicates the object classification. There is a strong relationship between collision likelihood and strength of the memory trace.

Participants. We recruited 50 participants from the Prolific research platform. The task took approximately 14 minutes, for which participants were compensated $3.50. We did not exclude any participants from our analyses.

Procedure. On each trial, participants would observe a Plinko scene with one of the objects highlighted, and would be asked to judge "How likely is the ball to collide with this object?" Participants indicated their response on a slide ranging from "not at all sure" to "very sure", which was again translated into a value between 0 and 100.

Stimuli. The scenes and highlighted objects were the same scenes and objects used in the critical memory trials. We counterbalanced the worlds and objects that participants saw in the exact same manner as experiment 1, such that each participant saw every scene once, made a collision judgment about a single object in each world, and made a collision judgment for each class of object four times.

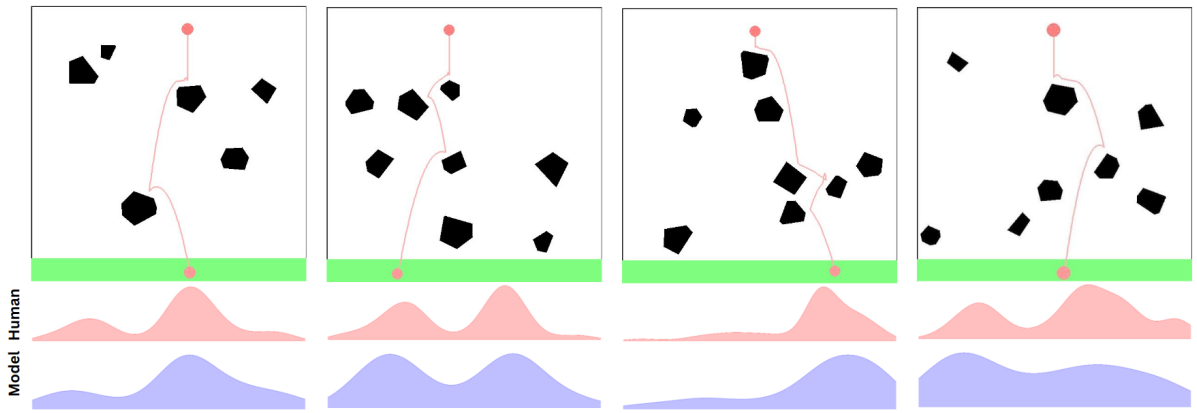

Figure 5: Four example trials and distributions of predictions for where the ball will end by participants (top) and the simulation model (bottom). Across trials, simulation captures the modes and general distribution of peoples' predictions, replicating the findings of Gerstenberg, Siegel, and Tenenbaum (2021).

Results. We find that these predictions of how likely the ball was to collide with an object act as a good proxy for relevance: the judgments of likelihood to contact from this experiment are well correlated with the memory for those same objects in Experiment 1 ($r = 0.81$, $p = 3.5 * 10^{-13}$, Fig. 4).

Measures of Relevance

Here we test the form of "relevance" that people use to decide which objects to include in our representations, as a way to constrain models of the cognitive processes that we use to construe the world. Specifically we consider two broad definitions of relevance: *simulation-based* relevance under which objects are included based on how often they would have *any* effect on simulation, and *value-based* relevance under which objects are included based on how much of an effect those objects have on the ultimate outcome of the system (e.g., where the ball ends up). The simulation-based relevance measure would better support the theory that we form representations "just in time," as this theory suggests that representations should be strengthened any time those objects will be contacted, regardless of how they change the trajectory of the ball. On the other hand, the value-based measure requires estimating the impact that an object on the outcome, and therefore better supports construal processes that learn to perform these estimates offline.

Simulation-based relevance. We measured simulation-based relevance of an object by instantiating a noisy simulation model, and counting the proportion of simulated paths that made contact with that object.

We relied on a modified version of the simulation model that was used to construct the stimuli, increasing the uncertainty in resolving collisions (Smith & Vul, 2013) in order to produce a better match between the distribution of the model's and human predictions of the ball's end location (correlation between average position for each trial: $r = 0.73$, see Fig. 5 for examples). We used this model to simulate 100 different ball trajectories, and calculated the number of times each object was contacted by the ball across all simulations. This metric correlated well with participants' average judgments of how likely the ball was to contact each object in Experiment 2 ($r = 0.96$), validating the model predictions. This simulation-based relevance explained participants' memory well ($r = 0.81$, Fig. 6, left), at around the same level as other participants' judgments of collision chance.

Value-based relevance. We instantiated the value-based relevance model as a measure of how much impact each object had on the end position of the ball. Determining impact requires comparing the outcome to alternative worlds, so we consider two transformations that are theorized to underlie how we create alternatives for counterfactual physical reasoning (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021): translations, where the object is shifted by a set amount, and deletions, in which the object is removed entirely from the world. The translations used to determine an alternative world were the same as were used for the memory test distractors (see Experiment 1: Stimuli).

In order to calculate the metric, we used the same noisy model as for the simulation-based relevance to produce a distribution of the ball's end location over 100 simulations, then applied the transformation and produced another distribution of 100 ball end locations. The relevance was calculated as the Wasserstein distance between these two distributions.

We find that both the translation and deletion models could explain some of the variability in participants' memory judgments (translation: $r = 0.59$, Fig. 6, middle; deletion: $r = 0.58$, Fig. 6, right), albeit not as well as the simulation-based metric.

Comparisons of relevance. While both measures of relevance explain participants' representations to some extent, we would expect some overlap due to correlations between the metrics: if an object is in the simulation path, then we should expect shifting or removing it will impact the trajectory, and if it is not in the path, then most of the time a shift or deletion will not have any effect. Indeed we do find that there is a strong correlation between the two metrics (simulation vs. translation: $r = 0.69$, simulation vs. deletion: $r = 0.74$), lead-

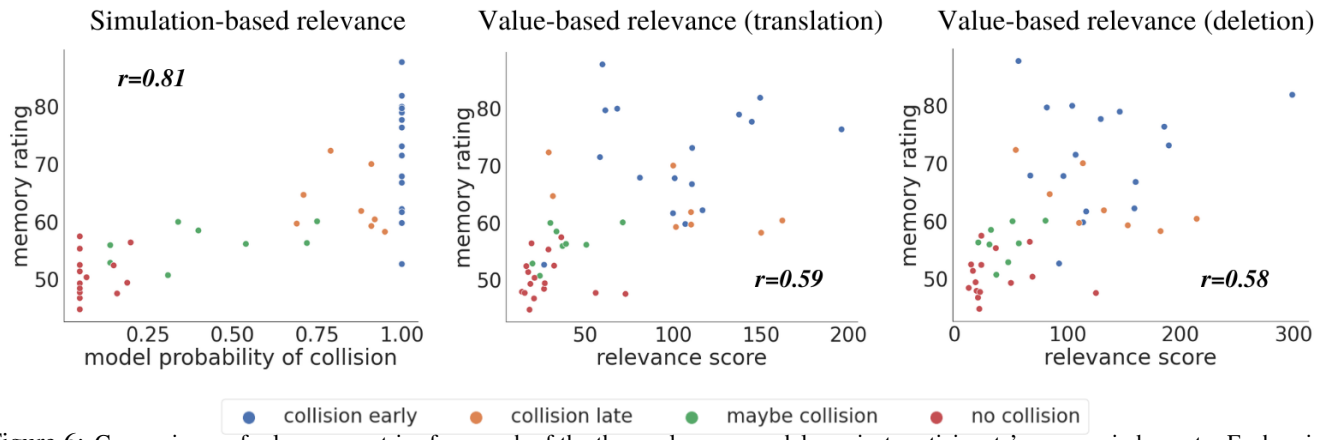

Figure 6: Comparisons of relevance metrics from each of the three relevance models against participants' memory judgments. Each point represents a different critical object; color indicates object classification. The simulation-based relevance metric explains memory for specific objects better than either value-based metric.

ing us to ask here whether they are explaining unique variance in participants' memory judgments.

Compared to a model that only uses simulation-based relevance to explain participants' memory judgments, adding either value-based relevance metric does not add any explanatory value ($p = 0.80$, $p = 0.52$ for translations and deletions, respectively). Conversely, adding simulation-based relevance on top of either of the value-based relevance metrics provides additional explanatory power ($p < 10^{-15}$ for both). This suggests that people rely on simulation-based relevance, providing evidence for "just in time" formation of representations.

Discussion and Future Directions

Our results suggest that people do not fully represent physical environments, but instead track the most relevant objects, perhaps by adding them "just in time" as they become relevant to simulation. Across a series of behavioral and model-based measures of relevancy, we find that humans construct object representations that closely match the probabilistic structure of the environment in a physical prediction task, such that recall for a particular object is highly tied to the probability of that object appearing in one's mental simulation.

We found that forward simulation measures such as the likelihood of collision were the best predictors of memory, as opposed to a more counterfactual measure that explicitly reasoned about the effects of the object change on the end state of the trajectories. However, here we only tested two counterfactual relevance measures, both measuring how a transformation applied to an object would impact the end position of the ball. Future work could investigate additional notions of counterfactual relevance, including more varied and systematic transformations such as small rotations and edge perturbations. Additionally, directly eliciting behavioral judgments of counterfactual relevance would potentially allow for more insight into the role that counterfactual reasoning plays in forming reduced representations.

While we were able to implicitly probe for information sampling and simulation via behavioral measures such as memory ratings and collision judgments, we are unable to fully disentangle the role of visual attention from rational representational compression. For instance, it is possible that the memory effects we observed fall out of the fact that people tend to look where they are simulating, and more relevant objects by definition lie closer to the simulated trajectory of the ball. A natural next step would therefore be to use eye-tracking and more controlled stimuli to investigate the joint roles of vision and memory in a more fine-grained way.

Finally, forming representations in more complex, real-world scenes will require extensions to the "just in time" framework. For instance, the current work primarily focused on how we represent static objects when there is one moving object. If there are more moving objects, they might collide with the critical object in our simulations (e.g., a ball flying in from the side), but determining whether this will happen (and thus if the other object is relevant) requires already representing and simulating that object's trajectory. Understanding how we choose what moving objects to represent requires further study, but could explain errors stemming from "partial simulation" (Bass et al., 2022). In addition, many scenes will have a number of potentially relevant objects that exceed working memory limits – such as block towers (Battaglia et al., 2013) – and prior work has suggested that we cannot represent each object individually (Ludwin-Peery et al., 2021). Future research should study not only how objects are added to our mental representations, but also how we might produce memory-constrained representations by, e.g., unloading objects from memory, or "grouping" multiple objects together.

The physical world is complex and filled with myriad numbers of objects, yet people can easily build a representation of the world that is constrained by our cognitive resources but still allows for effective predictions. Here we take a first step in explaining how the mind accomplishes this by showing that those representations are directly tied to their usefulness for simulation. In the future, we hope to explain how the mind decides what is useful in the first place, and how this representation is built up in tandem with physical simulation.

Acknowledgements

KAS and JBT were supported by National Science Foundation Science Technology Center Award CCF-1231216, NSF grant 2121009, and the DARPA Machine Common Sense program.

References

Ahuja, A., & Sheinberg, D. L. (2019). Behavioral and oculomotor evidence for visual simulation of object movement. *Journal of Vision*, *19*(6), 13. doi: 10.1167/19.6.13

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310. doi: 10.1073/pnas.1912341117

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2022). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413–424. doi: 10.1080/02643294.2022.2083950

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*, *127*(5), 891–917. doi: 10.1037/rev0000197

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2), 11. doi: 10.1167/19.2.11

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. doi: 10.1073/pnas.1306572110

Beller, A., Xu, Y., Linderman, S., & Gerstenberg, T. (2022). Looking into the past: Eye-tracking mental simulation in physical inference. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Bhui, R., Lai, L., & Gershman, S. J. (2021). Resource-rational decision making. *Current Opinion in Behavioral Sciences*, *41*, 15–21.

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. L. (2018). A resource-rational analysis of human planning. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (p. 6).

Crespi, S., Robino, C., Silva, O., & de'Sperati, C. (2012). Spotting expertise in the eyes: Billiards knowledge as revealed by gaze shifts in a dynamic visual prediction task. *Journal of Vision*, *12*(11), 30–30. doi: 10.1167/12.11.30

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

Emrich, S. M., Lockhart, H. A., & Al-Aidroos, N. (2017). Attention mediates the flexible allocation of visual working memory resources. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(7), 1454. (Publisher: US: American Psychological Association) doi: 10.1037/xhp0000398

Erev, I., & Barron, G. (2005). On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. *Psychological Review*, *112*(4), 912–931. doi: http://dx.doi.org/10.1037/0033-295X.112.4.912

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. doi: 10.1126/science.aac6076

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological review*, *128*(5), 936.

Gerstenberg, T., Siegel, M., & Tenenbaum, J. (2021). *What happened? Reconstructing the past through vision and sound.* PsyArXiv. doi: 10.31234/osf.io/tfjdk

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

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136.

Li, Y., Wang, Y., Boger, T., Smith, K., Gershman, S. J., & Ullman, T. (2022, February). *An Approximate Representation of Objects Underlies Physical Reasoning.* doi: 10.31234/osf.io/vebu5

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762–794. doi: http://dx.doi.org/10.1037/rev0000075

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*. doi: 10.1017/S0140525X1900061X

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021, June). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, *127*, 101396. doi: 10.1016/j.cogpsych.2021.101396

Musslick, S., Saxe, A., Hoskin, A., Reichman, D., & Cohen, J. D. (2020). On the rational boundedness of cognitive control: Shared versus separated representations. *PsyArXiv*.

Siegler, R. S. (1999). Strategic development. *Trends in Cognitive Sciences*, *3*(11), 430–435. doi: 10.1016/S1364-6613(99)01372-8

Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different Physical Intuitions Exist Between Tasks, Not Domains. *Computational Brain & Behavior*, *1*(2), 101–118. doi: 10.1007/s42113-018-0007-3

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

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017, September). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences*, *21*(9), 649–665. doi: 10.1016/j.tics.2017.05.012

Zhou, L., Smith, K., Tenenbaum, J., & Gerstenberg, T. (in press). Mental Jenga: A counterfactual simulation model of physical support. *Journal of Experimental Psychology: General*. doi: 10.31234/osf.io/4a5uh