Probabilistic simulation supports generalizable intuitive physics

Haoliang Wang[†1], **Khaled Jedoui**[2], **Rahul Venkatesh**[2], **Felix Binder**[1],
Josh Tenenbaum[3], **Judith Fan**[1,2], **Daniel Yamins**[2], **Kevin Smith**[3]
[1] University of California San Diego, [2] Stanford University, [3] Massachusetts Institute of Technology
[†]haw027@ucsd.edu

Abstract

How do people perform general-purpose physical reasoning across a variety of scenarios in everyday life? Across two studies with seven different physical scenarios, we asked participants to predict whether or where two objects will make contact. People achieved high accuracy and were highly consistent with each other in their predictions. We hypothesize that this robust generalization is a consequence of mental simulations of noisy physics. We designed an "intuitive physics engine" model to capture this generalizable simulation. We find that this model generalized in human-like ways to unseen stimuli and to a different query of predictions. We evaluated several state-of-the-art deep learning and scene feature models on the same task and found that they could not explain human predictions as well. This study provides evidence that human's robust generalization in physics predictions are supported by a probabilistic simulation model, and suggests the need for structure in learned dynamics models.

Keywords: intuitive physics; scene understanding; physical prediction; computational modeling

Introduction

Every day, we interact with the physical world in a variety of ways. We might start the morning by pouring cereal into a bowl, and later stably stacking that bowl with the rest of the dishes in the sink. Around the house, we might use a book to stabilize a wobbly chair, or throw trash into the trash can. Later with our kids we might build towers with blocks, or determine the right shot in a game of billiards. All of these scenarios require knowledge of many different principles of physics: from containment to stability to ballistic motion to collision dynamics. Yet we handle each of these tasks naturally, and often with little effort. But how are people able to do such general-purpose physical reasoning?

One hypothesis that has grown in prominence over the past decade is that we have a cognitive module that can perform general purpose, probabilistic physics simulation, often termed the *Intuitive Physics Engine* (Battaglia, Hamrick, & Tenenbaum, 2013; Ullman, Spelke, Battaglia, & Tenenbaum, 2017; Smith et al., in press). Under this hypothesis, general physics understanding arises because the simulation engine contains more primitive components for modeling the world – representations of objects and the forces they exert on each other, latent properties such as mass or elasticity that constrain how objects respond to forces, and key dynamic quantities such as momentum and events such as collisions – and combines them with uncertainty about the state of the world to reason probabilistically about a wide range of scenarios we might expect to encounter in day-to-day life. Thus, much like the physics engines that underlie many computer simulations, these building blocks of knowledge can be combined to model much more complex and combinatorial situations.

While this hypothesis has received quantitative support from many studies, a crucial aspect of it has never been explicitly tested. Prior studies that model human intuitive physics have typically focused on just one scenario at a time: e.g., how or whether a stack of objects might fall (Battaglia et al., 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Zhou, Smith, Tenenbaum, & Gerstenberg, 2023), how moving objects will bounce off each other and fixed obstacles (Smith & Vul, 2013; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Neupärtl, Tatai, & Rothkopf, 2021), or how liquid will pour (Bates, Yildirim, Tenenbaum, & Battaglia, 2019; Kubricht et al., 2016). While these models all use a physics engine at their core, across this research, modelers make different assumptions about the particulars of the physics engine, and fit different parameters to capture uncertainty about the state of the scene or how physical events resolve. This modeling approach risks overfitting to specific scenarios, and thus cannot answer the question of whether people have a general purpose physics simulator, or use different systems for different physical principles.

Indeed, another theory of human physical reasoning is that our judgments are based on inferences from past experience. This idea was first manifested in exemplar-based models and simple heuristics (Gilden & Proffitt, 1989; Nusseck, Lagarde, Bardy, Fleming, & Bülthoff, 2007; Proffitt, Kaiser, & Whelan, 1990; Sanborn, Mansinghka, & Griffiths, 2013): that people might base their judgments exclusively on combinations of features of the initial scene configuration without explicit reference to physical dynamics. Similar ideas have also been expressed by recent neural network models that learn to predict dynamics by watching videos. Proponents of this approach suggest that learning physics from raw data provides two benefits: these models can extract generalizable physical principles more flexibly than if the models were to rely on a fixed simulator, and can work directly from visual inputs in a way that physical simulation models on their own do not. A range of models have been proposed that express a spectrum of assumptions about what parts of physics should be learned, from those that attempt to jointly learn a scene representation and dynamics with few assumptions about the structure (Babaeizadeh et al., 2020), to models that assume the scene structure is known and try to learn only how objects interact (Han et al., 2022), and many in between. While these neural networks are often intended purely to advance an AI system's understanding of the physical world, they have been proposed as hypotheses for how infants learn physics (Piloto, Weinstein, Battaglia, & Botvinick, 2022), and have been used

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

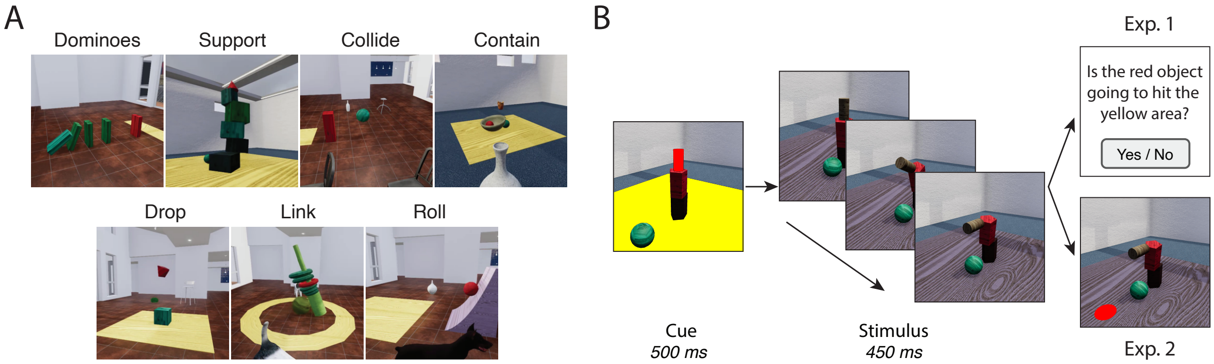

Figure 1: A: The seven different scenarios testing different physics principles. In each image, the object colored in red is the target object and the yellow area on the ground is the zone. B: Participants are cued with the target and zone objects, observe a short video, and predict either whether or where the target will contact the zone.

to predict both behavior and neural activity in monkeys performing physics prediciton tasks (Nayebi, Rajalingham, Jazayeri, & Yang, 2023).

In this paper, we test whether human physical predictions can be explained by approximate probabilistic inference in a single, general physics simulator across a wide range of everyday settings. We use an adapted version of the Physion dataset (Bear et al., 2021) which was designed to test arbitrary models' physics understanding in a variety of scenarios against both ground truth and human beliefs. We specifically test the *generalizability* of models: how well models can explain human predictions in scenarios that they have not been fitted or trained on. We show that an intuitive physics engine generalizes to these unseen scenarios in a human-like way, explaining human behavior only slightly worse than expected by the noise ceiling. We compare a variety of state-of-the-art deep learning models that encompass a range of assumptions about what is learned. Some jointly learn representations and dynamics with little structure, testing whether physics can be learned directly from video. Others learn to parse images into scene representations in a variety of ways – from few assumptions about scene structure to strong assumptions about 3D world structure – and learn dynamics on top of that representation with a recurrent network, in order to test how well these models produce representations that support learning physics. We also compare models that make heuristic predictions based on initial scene features. We find that an intuitive physics engine model captures human judgments and generalizes to unseen scenarios and novel tasks remarkably well, and far better than the deep learning and feature-based models tested. These results both support the mental simulation hypothesis as a generalizable mechanism for intuitive physical reasoning and point to the value of including stronger and more structured inductive biases into neural network models of intuitive physics.

Human experiments

To evaluate people's physical predictions across a wide range of scenarios, we adapted seven rigid body scenes from the Physion (Bear et al., 2021) dataset. These seven scenarios test a variety of physical concepts (Fig. 1A): chains of collisions (*dominoes*), the stability of a stack of objects (*support*), the particular ways collisions resolve (*collide*), whether one object can contain another (*contain*), how individual objects fall (*drop*), how a collection of objects can be knocked over (*link*), and how objects roll or slide down a slope (*roll*).

In each trial, there is a target object and zone. The goal is to predict either whether (Exp. 1) or where (Exp. 2) the target object will contact the zone at some point in the future.

Each scenario consists of 150 trials (1050 total), varying in scenario-specific configurations (see Bear et al. (2021) for details on their construction). Each trial consisted of a 450ms video in which the target object does not yet touch the zone. The trials were designed so that if the video had continued, in half of them the target would touch the zone within the next 2 seconds, but would never touch in the other half.

Experiment 1: Will it collide?

We first asked participants to make a binary judgement of whether they think the target object will contact the zone after watching a short video clip.

Participants 350 participants (50 per scenario; 198 female; all native English speakers) recruited from Prolific completed the experiment. Each participant was shown all 150 stimuli from a single scenario. Data from 33 participants were excluded for failing our preregistered inclusion attention checks. The experiment lasted approximately 15 minutes and participants were paid $3.50.

Task procedure The structure of our task is shown in Fig. 1B. Each trial began with a 500-1500ms fixation cross. Participants then saw the first frame of the video for 500ms with the target object and zone flashing a red and yellow overlay respectively, followed by the stimulus video for 450ms. Participants then saw a screen with buttons to indicate "YES" (the target would contact the zone) or "NO" (it would not). Before the main task, participants observed 10 familiarization trials for which the full movie was shown post-prediction.

Results We first examined how often participants' predictions of contact agreed with the simulation outcome from the

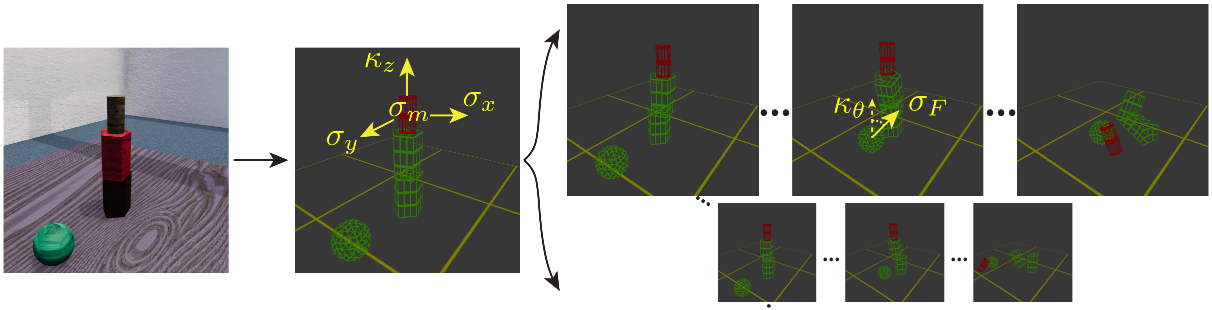

Figure 2: The Intuitive Physics Engine (IPE). The scene (left) is perceived to form an internal representation (middle), that includes perceptual (σ_x, σ_y, κ_z) and physical property (σ_m) uncertainty. The IPE then uses this representation to probabilistically simulate (right) how the world might unfold based on dynamic uncertainty (σ_F, κ_θ).

physics engine used to create the stimuli. We found that people achieved high accuracy (proportion correct = 0.81, 95% CI=[0.77, 0.84]) and their performance was substantially above chance across all seven scenarios ($t(6) = 16.49$, $p < 10^{-5}$). Participants' data also demonstrated variation in performance across scenarios, achieving the highest accuracy in Dominoes (0.84) and lowest in Roll (0.74).

Even though people made errors on some trials, these errors were consistent across participants (cross-trial bootstrapped split-half reliability=0.94, 95% CI=[0.92, 0.96], Fig. 3A). This pattern of high but imperfect accuracy and reliable errors is especially useful when comparing models with humans: to be a good explanation of how people make physical predictions across scenarios, a model should not only achieve high accuracy, but also err in the same ways as people do.

Experiment 2: Where will they touch?

Here we investigate more fine grained predictions by asking participants to indicate where they believe the target object will first contact the zone.

Participants A separate group of 245 participants (35 per scenario; 157 female; all native English speakers) recruited from Prolific completed the experiment. The experiment lasted approximately 16 minutes and paid \$3.75.

Task procedure The task procedure was identical to the "Will it collide" task except that participants were asked to place a circular disk where they believed the target would first contact the zone (Fig. 1B). For each scenario, the stimuli were the same as the previous experiment except that we filtered the 150 trials to only include trials where (a) the target object contacted the zone, and (b) this collision happened at a location that was unoccluded by other objects (Collide: 50 trials, Contain: 44, Dominoes: 68, Drop: 63, Link: 63, Roll: 65, Support: 48). After showing the first 450ms of the stimulus, the video froze on the final frame and participants used their cursor to position a disk on the target zone. Only the part of the disk overlapping the zone was displayed. When participants had placed the disk at their desired location, they clicked a "NEXT" button to register their prediction.

Results For each trial, we measured the center point of participants' disk placement positions as the 3D location in world coordinates. We excluded participants' placements that were off the zone by the disk radius (i.e. the disk had no overlap with the zone at all during the experiment, indicating that participants were not following the instructions or misclicked), accounting for about 5% of the data.

To assess how far off people are from the contact points given by the ground truth stimulus, we first calculated the Euclidean distance between the mean human predictions and the ground truth contact point for each trial, and averaged across trials. Because this metric is sensitive to the area of the target zone, for each trial we divided the distance by the standard deviation of participants' placements on that trial (analogous to d' in signal detection theory). We found similar patterns to the "will it" task: participants' predictions are significantly closer to ground truth than expected by chance, and this is true for every physical scenario (mean normalized distance=1.39, 95% CI=[1.23, 1.51], $t(6) = 43.25$, $p < 10^{-8}$). Furthermore, we calculated the split half distance between participants (the average distance between the mean predictions of evenly splitting participants into two random groups) and found that they are highly consistent with each other (mean distance=0.50, 95% CI=[0.38, 0.71], Fig. 3B).

The Intuitive Physics Engine

The two experiments described previously demonstrate that across a wide range of physical scenarios, people make good predictions but are also biased in systematic ways. We argue that these predictions can be characterized using a noisy physics engine that runs probabilistic simulations – that the characteristic patterns of errors and biases we observe in participants' data can be mostly explained by uncertainty about the state of the scene after watching the video plus a noisy, approximately correct simulator that transforms those initial states into a distribution over outcomes. We formalize this hypothesis in an Intuitive Physics Engine (IPE) model.

The architecture of the IPE

To model people's prediction in naturalistic 3D environments, we used Unity3D as the underlying physics engine and customized to add noise to model sources of uncertainty in humans. Following Smith and Vul (2013), we considered uncertainty along three different axes (see Fig. 2):

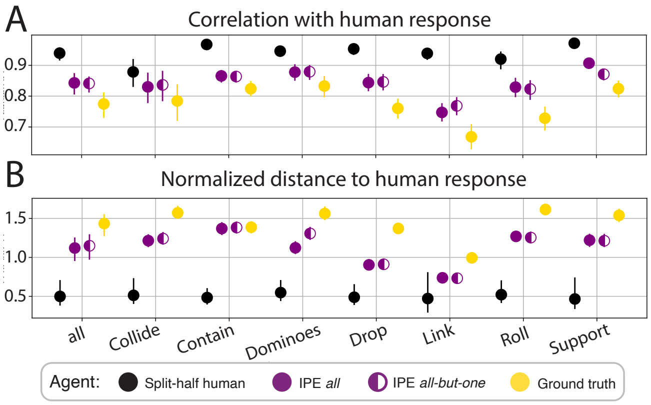

Figure 3: Comparison between the IPE and ground truth to human response on individual scenarios: Exp. 1 (A), Exp. 2 (B). Across all seven scenarios, the IPE better captures human response patterns than ground truth, and loses almost no predictive power across scenarios in the "all-but" fitting regime. Error bars are 95% CIs.

Perceptual uncertainty: We modeled people's uncertainty in visual perception by adding noise to the initial positions and rotations of the objects. The starting position of each of the objects was perturbed around the true position by two-dimensional Gaussian noise parameterized by standard deviation σ_x, σ_y; and the rotation by von Mises noise around the z axis parameterized by concentration κ_z.[1]

Physical property uncertainty: We capture people's uncertainty about physical properties that vary across objects but are not directly observable – the mass of different objects – by adding Gaussian noise to the true mass, parameterized by standard deviation σ_m, truncated at zero.

Dynamic uncertainty: We considered people's uncertainty about how collision will resolve, by perturbing the resultant collision impulse force's magnitude by Gaussian noise around its true value with the standard deviation σ_F, and direction by a spherical von Mises distribution centered on the true angle of the impulse with a concentration parameter κ_θ.

Fitting model parameters

In order to determine the set of noise parameters that best describes human behavior, we fit the six parameters defined above to participants' data on the "will it" task. Because the "where" task requires no additional modeling assumptions, we can use the same model to compare to participants' data on this task, and thus can treat model performance on that task as generalization to a separate task.

We fit parameters by simulating the scenes for 2.5s with the noisy IPE 20 times for each scene. We measured the RMSE between the proportion of IPE runs that predict contact for each trial, and the proportion of participants that do. We minimized this RMSE using the HyperOpt package (Bergstra, Yamins, & Cox, 2013).

We used two regimes for fitting. In the "all scenarios" regime, we fit the IPE to 20% of trials from each scenario (210 trials total). We then assessed performance on the 80% of trials the model had not been fit on (840 trials), testing generalization to new trials. In the "all-but-one" regime, we fit seven separate IPE models: each fit on the trials from six of the seven scenarios, then assessed performance of each of those models on the unseen scenario. Overall model performance was calculated by averaging over the performance of each model on its held-out scenario. This regime tests even stronger generalization: whether the uncertainty measured in separate scenarios can explain human predictions in scenarios uninvolved in the fitting (Wang, Allen, Vul, & Fan, 2022).

Model Results

We first test whether a single noisy simulator can explain a range of human judgments by evaluating the IPE's predictions against humans' on both the "will it" and the "where" task. We then assess how well a set of state-of-the-art deep learning networks explain human predictions.

The IPE is physical-domain-general

Will it contact? We find that a single parameterization of the IPE can explain human judgments across scenarios. Using the "all" fitting regime, the IPE achieves human-level performance across all seven scenarios on the test set (mean accuracy=0.83, 95% CI=[0.79, 0.86], Fig. 4A), and more importantly also has high correlation with human responses (mean correlation=0.87, 95% CI=[0.83, 0.90], Fig. 3A), only slightly worse than could be expected by the human noise ceiling. We also compare against how well participants would be fit by assuming perfectly accurate predictions, and find that the IPE correlates with human predictions better ($t(13)$=2.42, $p = 0.01$, Fig. 3A). This indicates that the IPE not only captures overall human performance, but also makes similar predictions on individual trials. Importantly, the IPE explains predictions better than the ground truth answers, suggesting that the uncertainty inherent in the model leads to uncertainty in outcomes that produce human-like errors across scenarios.

As a stronger test of generalization, we assess how well the IPE explains human data in the "all-but-one" fitting regime, where the model is assessed on scenarios it has not observed during parameter fitting. The IPE maintained high accuracy (mean accuracy=0.81, 95% CI=[0.76, 0.86]) and correlation with human responses (mean correlation=0.87, 95% CI=[0.84, 0.89]), and was nearly the same when it had access to trials from all scenarios – a pattern that held across all scenarios (Fig. 3A). Thus uncertainty about physical properties can be assessed in one set of scenarios and extrapolated to separate scenarios without noticeably affecting performance.

Where will it contact? To evaluate the IPE against humans on precise location predictions, we reused the noise parameters fit on the "will it" task and extracted the contact location information between the target object and zone. The pattern of results is similar to those for the "will it" task: the

[1] Following Battaglia et al. (2013), we only consider position and rotation uncertainty along a plane because most objects are resting on the ground or another object, and so uncertainty along the z-axis would cause objects to either float or interpenetrate.

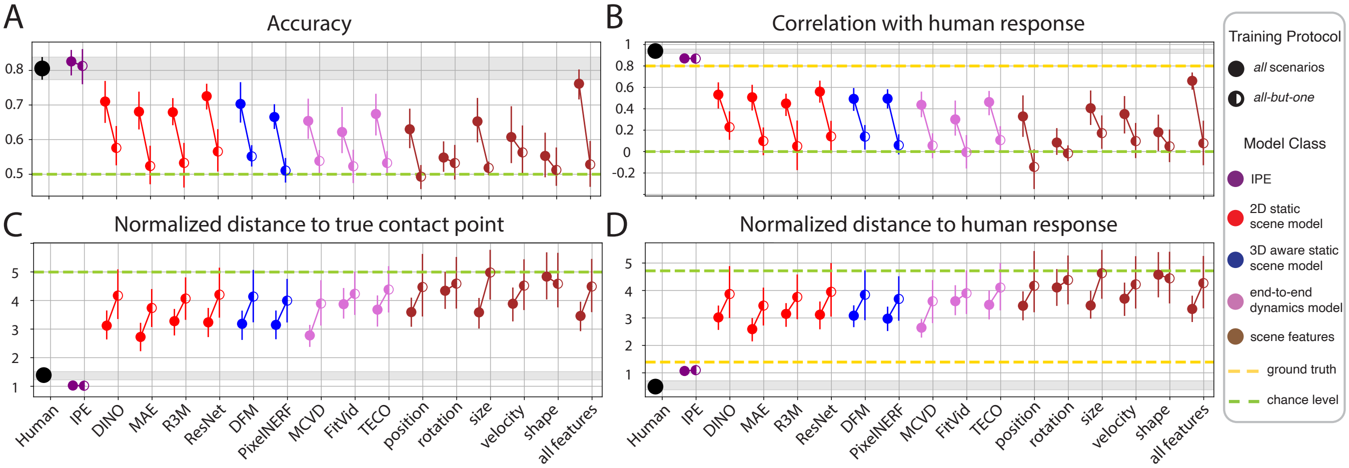

Figure 4: Performance of people vs. models on the two experiments. Exp 1: A. proportion of correct contact judgments, B. correlation to human responses. Exp 2: C. normalized distance to the true contact point, D. normalized distance to average human prediction. The IPE performs near human-levels across all task metrics, maintained when generalizing across scenarios. All deep learning and feature-based models perform worse than the IPE, and their performance suffers when generalizing in the "all-but-one" training protocol.

IPE's distance to the true contact remained the same across fitting regimes (all: 1.03, 95% CI=[0.88, 1.21], all-but-one: 1.02, 95% CI=[0.88, 1.18]; Fig. 4C). The average normalized distance between the model's predictions and human placements is 1.07, 95% CI=[0.89, 1.21], significantly less than the distance between people and ground truth ($t(13)$=2.64, $p = 0.01$) and at the same time did not change when generalizing across scenarios (normalized distance=1.10, 95% CI=[0.91, 1.25], Fig. 3B). Note that the "all-but-one" fitting regime is an extremely strong test of generalization: the model must generalize across participants, physical scenarios, and even the type of prediction.

However, the IPE does capture human performance slightly worse on the "where" task than the "will it" task. This could be due to the aforementioned strong generalization hindering performance, because, e.g., one set of participants have different amounts of uncertainty than the other, or because simply judging whether contact will occur is a coarser measure than judging where contact would occur, and so parameter estimates should be less precise.

Comparing to deep learning and feature models

While we have shown that the IPE can explain human predictions across scenarios, another theory suggests human-like physics understanding can arise from less structured learning. Thus, in this section, we aim to test the generalizability of state-of-the-art deep learning models as well as models that make physical predictions based on scene features, and compare their predictions on the same stimuli to humans and IPE.

We selected state-of-the-art models from three representative model architecture classes. These models were either pretrained or finetuned on the Physion dataset. **(1)** We assessed a set of 2D static scene understanding models with an LSTM trained on Physion scenes to predict next-frame dynamics from scene representations with various amounts of structure (*DINO*, Oquab et al., 2023; *MAE*, He et al., 2021; *R3M*, Nair, Rajeswaran, Kumar, Finn, & Gupta, 2022; and *ResNet* He, Zhang, Ren, & Sun, 2016). This assesses whether the scene representations learned by these models support efficient learning of dynamics. **(2)** We assessed 3D scene understanding models with the same LSTM training on their scene representations (*DFM*, Tewari et al., 2023, and *PixelNERF*, Yu, Ye, Tancik, & Kanazawa, 2021). This assesses whether richer, 3D-aware scene representations might support better prediction. **(3)** We assess end-to-end dynamics models pretrained on Physion scenes (*MCVD*, Voleti, Jolicoeur-Martineau, & Pal, 2022; *FitVid*, Babaeizadeh et al., 2020; and *TECO*, Yan, Hafner, James, & Abbeel, 2022). These models asses whether human-like physics knowledge could be learned in an unstructured manner from video.

In order to compare deep learning models to humans on the "will it" task, for each model, we first extracted features by showing the human stimulus (450ms) and concatenated them with the "simulated" features output by the model's dynamics predictor. To get a binary output from the models, we then froze the parameters of the model and fit a logistic regression on the features. The parameters for the logistic regression were fit on a separate set of stimulus provided by the Physion dataset, with the ground truth object contact labels acting as supervision. We evaluated these models on the same unseen 840 experimental trials that the IPE was evaluated on. As seen in Fig. 4 AB, none of the deep learning models reached human levels of accuracy, and they did not correlate with human predictions as well as the IPE. In the "all-but-one" regime we trained the deep models on six out of the seven scenarios and tested them on the held-out scenario, and found that performance dropped noticeably across the two tasks ($p < 10^{-3}$ for all models for both accuracy and correlation), with many of the models only marginally exceeding chance levels.

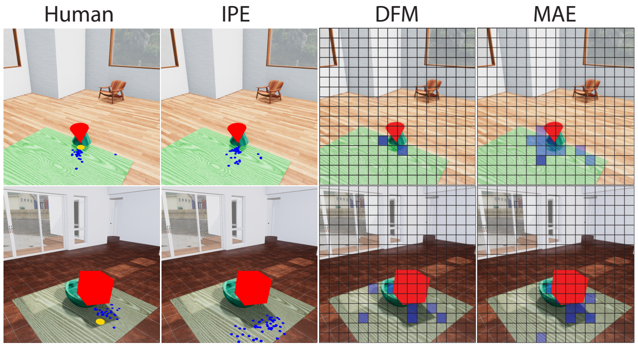

Figure 5: Two example "where" trials showing people's and models' predictions. The target object is highlighted in red, predictions are colored in blue. The ground truth contact point is colored in gold in the human panel.

Next, we evaluated the deep models' predictions of where it believed contact would occur. Unlike the IPE, all the deep models compute on 2D images rather than 3D world coordinates, so we used logistic regression on the same model features as before to output a prediction probability distribution on a 16×16 grid over the image and then projected the center of each cell in the grid to 3D world coordinates (see Fig. 5).[2]

In order to compare between humans, the IPE, and deep learning models, we needed to align their predictions. First, we transformed human and IPE predictions by translating their predictions on world coordinates back into 2D image coordinates, similarly binning them into 16 × 16 grids, approximating the prediction using center of the cell and then projecting the center back to 3D world coordinates. Second, because participants and the IPE were only allowed to make predictions on the zone area, we re-normalized the prediction probability distribution from the deep learning models to be only on the zone. We then measured probability-weighted distance between the grid center points on world coordinates as a metric for prediction distance.[3] As shown in Fig. 4 CD, the deep learning models did not capture human performance as well as the IPE, and always had decreased predictivity in the "all-but-one" regime, though here most performed above chance, providing evidence that they had learned something about the physics of these scenes as a whole.

We also evaluated feature-based models on the same tasks, using objects' position, rotation, size, shape and velocity at the stopping frame (i.e. 450ms) individually as features, as well as a combination of all five, and then fit a linear readout model in the same way we did for deep learning models for each feature (including fitting to the 16x16 grid and normalizing for the "where" task). The combination of all scene features could predict whether contact would occur relatively well in the "all" training regime, but these features could not generalize to unseen scenarios in the "all-but-one" regime (Fig. 4 AB). The scene features also poorly predicted "where" the objects would contact, with performance falling to close to chance in the "all-but-one" scenario (Fig. 4 CD).

Discussion

In this paper, we tested the hypothesis that human physical predictions can be explained by approximate probabilistic inference in a single, general physics simulator across a wide range of everyday settings. Across two experiments, we found that a physics engine that runs probabilistic simulations generalized to unseen stimuli in human-like ways, but a set of state-of-the-art deep learning models and feature-based models do not yet reach that level.

One major point of difference between the IPE and the deep learning models we tested here is the input encoding: the IPE takes 3D information of the scene as inputs whereas the deep learning models compute on pixels. Learning directly from pixels can allow for greater flexibility in the representation of scenes and dynamics, but imposes the challenge of learning to extract scene information rather than that information being provided. In this case, however, it appears these representations do not support longer term predictions in untrained scenarios that require understanding the physics of the world. In future work, we will evaluate a broader set of deep learning models that impose greater structure on the learning of physics – e.g., graph neural networks that work from scene representations and explicitly parse the world into objects and their relations (Mrowca et al., 2018; Li, Wu, Tedrake, Tenenbaum, & Torralba, 2018; Han et al., 2022; Battaglia, Pascanu, Lai, Jimenez Rezende, & Kavukcuoglu, 2016; Allen et al., 2022) – as well as noisy simulation models that rely on scene parsing models to provide information about the world (Wu, Lu, Kohli, Freeman, & Tenenbaum, 2017). Systematic testing of broader sets of models can help inform us what additional structure is required to develop more human-like understandings of the physical world.

While the IPE model predicts human response patterns well, it is still below the noise ceiling. This is a pattern found in many studies going back to Battaglia et al. (2013), and is likely due to the fact that humans cognitive simulations are not exactly the same as computer physics engines, but instead have different implementations and additional simplifications (Bass, Smith, Bonawitz, & Ullman, 2021; Chen, Allen, Cheyette, Tenenbaum, & Smith, 2023; Li et al., 2023). Further research into the structure of human physical representations and simulations will be required to close this gap.

The physical world is complex and open-ended, yet we easily reason about a wide range of scenarios that we might encounter in everyday life. The current study suggests that this robust generalization behavior often comes from having a generalizable mental model of the physical world and the ability to continuously simulate forward about how the world will unfold. In the long term, such studies will help us to understand and implement the computational mechanisms needed for the deep learning models to be more human-like.

[2] We found empirically that structuring our problem as a 16×16 grid classification task improved the readout training performance compared to a position regression task, while also guaranteeing fine-grained predictions.

[3] We also considered Wasserstein distance over grid distributions, and found qualitatively similar patterns of results.

Acknowledgments

We acknowledge Simons Foundation grant (543061), NSF CAREER grant (1844724), ONR grant (S5122), ONR MURI 00010802, and ONR MURI S5847, awarded to D.L.K.Y.; ONR Science of Autonomy award and Stanford HAI Hoffman-Yee grant awarded to J.E.F.; the DARPA Machine Common Sense grant awarded to J.B.T.; and NSF grant (2121009) awarded to J.B.T. and K.A.S.. We also thank the NVIDIA corporation and the Google TPU Research Cloud group for hardware grants.

References

Allen, K. R., Rubanova, Y., Lopez-Guevara, T., Whitney, W., Sanchez-Gonzalez, A., Battaglia, P. W., & Pfaff, T. (2022). *Learning rigid dynamics with face interaction graph networks* (No. arXiv:2212.03574). arXiv.

Babaeizadeh, M., Saffar, M. T., Nair, S., Levine, S., Finn, C., & Erhan, D. (2020). Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413–424.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLOS Computational Biology*, *15*(7), e1007210. doi: 10.1371/journal.pcbi.1007210

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.

Battaglia, P. W., Pascanu, R., Lai, M., Jimenez Rezende, D., & Kavukcuoglu, K. (2016). Interaction Networks for Learning about Objects, Relations and Physics. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29* (pp. 4502–4510). Curran Associates, Inc.

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., … others (2021). Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning* (pp. 115–123).

Chen, T., Allen, K. R., Cheyette, S. J., Tenenbaum, J., & Smith, K. A. (2023). " just in time" representations for mental simulation in intuittive physics. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2021). A counterfactual simulation model of causal judgment for physical events. *Psychological Review*. doi: 10.1037/rev0000281

Gilden, D. L., & Proffitt, D. R. (1989). Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 372.

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.

Han, J., Huang, W., Ma, H., Li, J., Tenenbaum, J., & Gan, C. (2022). Learning physical dynamics with subequivariant graph neural networks. *Advances in Neural Information Processing Systems*, *35*, 26256–26268.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked autoencoders are scalable vision learners. *arXiv:2111.06377*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Kubricht, J. R., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., & Lu, H. (2016). Probabilistic Simulation Predicts Human Performance on Viscous Fluid-Pouring Problem. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society.*

Li, Y., Wang, Y., Boger, T., Smith, K. A., Gershman, S. J., & Ullman, T. D. (2023). An approximate representation of objects underlies physical reasoning. *Journal of Experimental Psychology: General*.

Li, Y., Wu, J., Tedrake, R., Tenenbaum, J. B., & Torralba, A. (2018). Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*.

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fei-Fei, L. F., Tenenbaum, J., & Yamins, D. L. (2018). Flexible neural representation for physics prediction. *Advances in neural information processing systems*, *31*.

Nair, S., Rajeswaran, A., Kumar, V., Finn, C., & Gupta, A. (2022). R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*.

Nayebi, A., Rajalingham, R., Jazayeri, M., & Yang, G. R. (2023). *Neural Foundations of Mental Simulation: Future Prediction of Latent Representations on Dynamic Scenes* (No. arXiv:2305.11772). arXiv.

Neupärtl, N., Tatai, F., & Rothkopf, C. A. (2021). Naturalistic embodied interactions elicit intuitive physical behaviour in accordance with Newtonian physics. *Cognitive Neuropsychology*, 1–15. doi: 10.1080/02643294.2021.2008890

Nusseck, M., Lagarde, J., Bardy, B., Fleming, R., & Bülthoff, H. H. (2007). Perception and prediction of simple object interactions. In *Proceedings of the 4th symposium on applied perception in graphics and visualization* (pp. 27–34).

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., … Bojanowski, P. (2023). *Dinov2: Learning robust visual features without supervision.*

Piloto, L. S., Weinstein, A., Battaglia, P. W., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Be-*

haviour, 1–11. doi: 10.1038/s41562-022-01394-8

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

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Smith, K. A., Hamrick, J. B., Sanborn, A. N., Battaglia, P. W., Gerstenberg, T., Ullman, T. D., & Tenenbaum, J. B. (in press). Probabilistic models of physical reasoning. In *Bayesian models of cognition: Reverse-engineering the mind.* MIT Press.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Tewari, A., Yin, T., Cazenavette, G., Rezchikov, S., Tenenbaum, J. B., Durand, F., . . . Sitzmann, V. (2023). Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *NeurIPS*.

Ullman, T. D., Spelke, E., Battaglia, P. W., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, *21*(9), 649–665.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82. doi: 10.1016/j.cogpsych.2017.05.006

Voleti, V., Jolicoeur-Martineau, A., & Pal, C. (2022). Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In *(neurips) advances in neural information processing systems.* Retrieved from `https://arxiv.org/abs/2205.09853`

Wang, H., Allen, K. R., Vul, E., & Fan, J. E. (2022). Generalizing physical prediction by composing forces and objects. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Wu, J., Lu, E., Kohli, P., Freeman, W. T., & Tenenbaum, J. B. (2017). Learning to See Physics via Visual De-animation. In *Neural Information Processing Systems* (p. 12).

Yan, W., Hafner, D., James, S., & Abbeel, P. (2022). Temporally consistent transformers for video generation. *arXiv preprint arXiv:2210.02396*.

Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021). pixelNeRF: Neural radiance fields from one or few images. In *Cvpr.*

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023). Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*.