Neurocomputational Modeling of Human Physical Scene Understanding

Ilker Yildirim* (ilkery@mit.edu)[1,3] **Kevin A Smith* (k2smith@mit.edu)**[1,3] **Mario Belledonne* (belledon@mit.edu)**[1,3]
Jiajun Wu (jiajunwu@mit.edu)[2,3] **Joshua B Tenenbaum (jbt@mit.edu)**[1,2,3]

* Indicates equal contribution.

[1] Dept. of Brain & Cognitive Sciences, MIT [2] Computer Science and Artificial Intelligence Laboratory, MIT

[3] Center for Brains, Minds, and Machines, MIT Cambridge, MA 02139

Abstract

Human scene understanding involves not just localizing objects, but also inferring latent attributes that affect how the scene might unfold, such as the masses of objects within the scene. These attributes can sometimes only be inferred from the dynamics of a scene, but people can flexibly integrate this information to update their inferences. Here we propose a neurally plausible *Efficient Physical Inference* model that can generate and update inferences from videos. This model makes inferences over the inputs to a generative model of physics and graphics, using an LSTM based recognition network to efficiently approximate rational probabilistic conditioning. We find that this model not only rapidly and accurately recovers latent object information, but also that its inferences evolve with more information in a way similar to human judgments. The model provides a testable hypothesis about the population-level activity in brain regions underlying physical reasoning.

Keywords: intuitive physics, recurrent recognition networks, probabilistic simulation engines

Introduction

Scene understanding is not only about recognizing what is where, but seeing the physics of a scene. From a glance at the two cuboids in Fig. 1 (left), we expect the metal block on the right to be heavier than the wooden block on the left. There are many cases, however, where our visual estimation of object properties is underdetermined or even misleading (*e.g.*, the block on the right could be made of styrofoam but covered by a thin metallic sheet). Humans can infer physical object properties not only from static appearances, but from dynamic scenes: seeing the wood cuboid bump into the 'metal' one (Fig. 1, center) and cause it to move rather than stop the wood block's motion (Fig. 1, right), we realize our visual estimate of the objects' masses should be revised.

Studies of human physical inferences suggest that this process is supported by an "intuitive physics engine" (Battaglia, Hamrick, & Tenenbaum, 2013), which, similar to a video game engine, allows us to simulate possible ways the world will unfold. According to this framework, physical scene understanding amounts to "analysis-by-synthesis": setting and adjusting the initial scene configuration (*e.g.*, weights of objects) so that our simulations match our observations. These studies inspired several works in AI including attempts to build "neural" physics engines (Battaglia, Pascanu, Lai, Rezende, et al., 2016) and jointly modeling system dynamics and visual inputs (Wu, Lu, Kohli, Freeman, & Tenenbaum, 2017). Despite this progress, two key questions remain:

1. **Dynamic updates:** How are objects' property estimates updated dynamically as we continuously gather more information from the world?

Figure 1: A 'surprising' collision. If we observe a wooden block on a ramp with an iron block at the bottom (*left*), we expect the iron to stop the wood upon collision (*center*). Seeing the wooden block launch the iron one (*right*), lets us update our beliefs about the actual masses of the two blocks.

2. **Mappability:** How are such updating processes computed in the brain?

Here, we present a new computational account of human intuitive physical scene understanding. Our core idea is to formulate the problem of physical scene understanding as *efficient inference* in a generative model. The generative model is a probabilistic program, wrapping a physics engine to realistically animate a world of objects, and a graphics engine to render the evolving world states. Latent variables in this generative model are objects' substances (masses, frictions), geometries (shapes), and kinematic properties (positions, velocities, rotations, collisions).

We implement efficient inference with a recognition network where the network architecture follows the causal structure of the generative model. Given an unfolding video as input, the network's goal is to produce and dynamically update point estimates of the latent variables through time. We use an LSTM to integrate inference over time. At each time step, the recognition network updates the state of a single object in the scene while encoding the visual input using attention to emphasize the relevant evidence. This joint attention mechanism, or the binding parameters, allows the model to link visual evidence to internal object states. We name our model the Efficient Physical Inference network (EPI).

We evaluate EPI in a basic yet physically rich dynamical scenario: an object on the ramp slides and collides with another stationary object on the ground. We evaluate EPI as an inference algorithm by comparing it to a standard inference procedure for non-linear dynamical systems: an idealized (*i.e.* non image-computable) sequential Monte Carlo (SMC) algorithm with limited computational resources (small numbers of particles). We also compare both algorithms to an ideal observer model with no resource constraints. We find that EPI's inferences approximate that of the SMC while being faster.

Although we are not making direct contact to neural data in this study, we note that EPI is composed of neurally plausible components (*e.g.*, feedforward networks and RNNs). Each of its layers provides a testable computational hypothesis about the neural computation underlying physical reasoning regions

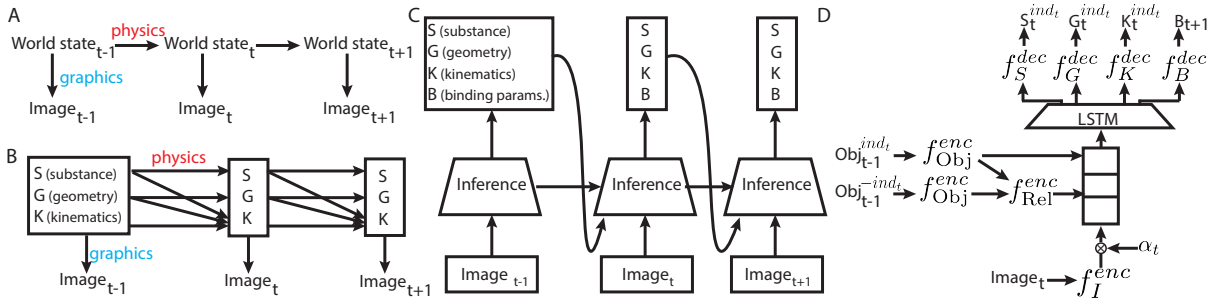

Figure 2: (A) Schematic of a generative process formalizing formation of dynamic scenes. (B) Schematic of our generative model. Latent variables are substance, geometry, and kinematics properties of objects in the sensory environment. A physics engine evolves the state of the objects in time. A graphics engine renders each state. (C) Schematic of the EPI recognition network. Its architecture is designed to match the causal structure in the generative model. (D) Details of the recognition network.

in the primate brain (Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016; Sliwa & Freiwald, 2017). Contrast this to SMC, where it is less clear how to map a software physics engine or the proposal-likelihood evaluation cycles to neural computation.

Our main results show EPI's ability to capture the temporal dynamics of human subjects' mass judgments across a broad range of stimulus conditions. These include conditions where object appearance and physical properties are intentionally incongruent, *e.g.*, a light wooden block with the appearance of an iron block. We show that the EPI network can generalize to such mismatch stimuli just as humans do, and provides a highly accurate quantitative account of human dynamic belief updating behavior. We also find the approximate inference schemes, both EPI and SMC, account for the human data better than the ideal observer model.

Efficient Physical Inference (EPI) Network

Generative model: Dynamic scenes can be formalized using a generative process of physics and graphics (Fig. 2A). We present an instantiation of this generative model in the ramp scenario (Fig. 2B). For object ind in the scene, latent variables consist of the object's state including its geometry G^{ind}, substance S^{ind}, and kinematic K^{ind} properties, collectively denoted as $\text{Obj}^{ind} = \{S^{ind}, G^{ind}, K^{ind}\}$.

Substance properties S include mass and friction. Each object is randomly assigned to a substance type (Iron, Brick & Wood). Density and friction are sampled from mixture distributions with two components: a normal distribution with the substance type's default density and friction parameters as its mean, and a uniform distribution over entire ranges. The mass of an object is then the product of its density and volume.

Geometry properties G include a categorical shape type (Block & Puck) and continuous height, width, and depth parameters. Kinematic properties K include positions, velocities, rotations, and collisions across time. The kinematic state of an object evolves using a 3D physics engine $K_t = \pi(S_{t-1}, G_{t-1}, K_{t-1})$. A graphics engine is used to render images at each time step, $I_t = \gamma(\text{Obj}^0, \text{Obj}^1, A)$ where A denotes fixed rendering parameters such as viewpoint and lighting.

Given a video, physical scene understanding can be cast as inference in this generative model.

$$\Pr(\text{Obj}^0, \text{Obj}^1 | I_{0:T}, \pi(\cdot), \gamma(\cdot)) \propto$$
$$\Pr(I_{0:T} | \text{Obj}^0, \text{Obj}^1, \pi(\cdot), \gamma(\cdot)) \Pr(\text{Obj}^0, \text{Obj}^1) \delta_\pi \delta_\gamma \quad (1)$$

Traditionally sampling based approaches have been the inference method for such intractable posteriors. We present two such methods with different computational resource limitations before presenting the EPI recognition network.

Inference using sampling based approaches: These inference schemes sample from an idealistic form of Eq. 1: They assume ground truth geometry G and initial kinematics state K_0; given noisy positions and velocities as observations, they estimate the substance properties, $\Pr(S^0, S^1 | K^0_{0:T}, K^1_{0:T})$.

The first, an *Ideal Observer* (IO-MH), represents how a resource-unconstrained Bayesian observer would perform inference in this generative model. It is an MCMC algorithm with Metropolis-Hastings updates for S^0 and S^1. The second, a *Sequential Rational Process* (SRP) model is designed to dynamically update inferences about S^0 and S^1 as more of K^0 and K^1 is observed through time. It is implemented as a particle filter (an instance of SMC) with resampling and with occasional replacement from the prior for avoiding degeneracy.

EPI network

Network architecture: The EPI network aims to compile sequential inference in the generative model in a recurrent latent variable recognition network. The model's overall architecture is dictated by the generative model with video frames as inputs and latent variables as targets (Fig. 2C).

The model consists of four components (Fig. 2D). First is encoding the images, which takes as input a sequence of images (a video; $I_{0:T}$). At each time step t, the model encodes image I_t using a combination of f^{enc}_I, the top convolutional layer activations (TCL) of the Imagenet pre-trained Alexnet, and a 13×13 soft-max attention map for a weighted sum of the TCL activity, reducing it from a tensor of size $256 \times 13 \times 13$ to a vector of size 256.

Second is the encoding of objects and their relations following the formulations in object-based neural physics engines (*e.g.* Battaglia et al., 2016). We encode the state of the object to be updated, Obj^{ind} using f^{enc}_{Obj}, a multilayer perceptron (MLP). We encode their relationship using f^{enc}_{Rel}, an MLP, as $f^{enc}_{\text{Rel}}(f^{enc}_{\text{Obj}}(\text{Obj}^{ind}), f^{enc}_{\text{Obj}}(\text{Obj}^{-ind}))$.

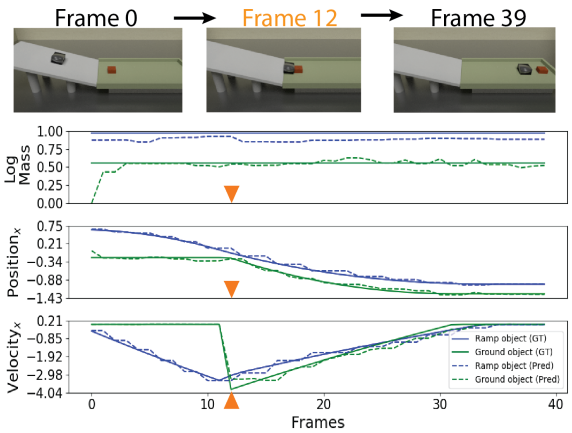

Figure 3: EPI inference trace. Changes in attributes over time for the ramp object (blue) and ground object (green), with model inferences (dashed lines). The orange triangle denotes the frame of collision.

Table 1: Model Evaluations

Model	Log-likelihood	Time (s)
EPI	$-6.33(5.6)$	$0.21(0.11)$
SRP	$-2.99(2.67)$	$0.36(0.05)$
Ideal observer (MH)	$-1.04(0.06)$	$120(49.5)$

Third, the model recurrently updates its state using an LSTM which takes as input concatenation of encoded image and encoded object and its relation. Fourth, the model updates the state of object ind, $\{S_t^{ind}, G_t^{ind}, K_t^{ind}\}$, using MLP decoders $f_S^{dec}, f_G^{dec}, f_K^{dec}$. These decoders takes as input the output of the LSTM. At each time step, the network also predicts the binding parameters, B_{t+1} using f_B^{dec}. This decoder consists of two independent MLPs, each taking as input the output of the LSTM, and predicting the binding parameters: which object's state to update at the next time step, ind_{t+1}, and where in the image to attend for its relevant evidence, α_{t+1}.

Objective functions: We minimize two objectives. First, we minimize the distance between predicted and true latent variables for each object at each time point, $d_{t,0}$ and $d_{t,1}$, using a smooth L1 loss. Second, we minimize a simple cross-entropy loss that allows the model to learn to predict which object's state it should update based on the LSTM's output without having $d_{t,0}$ and $d_{t,1}$ available during test time. We formulate this objective as $L_{ind} = \sum_{t=0}^{T} \sum ind_t \log(\hat{ind}_t)$, where ind_t is one-hot encoding of $\text{argmax}_i\{d_{t,i}\}$ and $\hat{ind}_t$ is an output of f_B^{dec}.

Training data: The model is trained using stochastic gradient descent based on samples drawn from the generative model in the style of Helmholtz machines without requiring labeled data. An example trace on a test item is shown in Fig. 3.

Speed and accuracy comparisons: We compared models' wall-clock execution time and their average log-likelihood scores of predicting ground-truth positions and velocities on a total of 168 trials (used in the behavioral experiment). We find EPI is on the accuracy-efficiency trade-off frontier, approaching the accuracy of IO-MH while running faster than SRP (Table 1).

Testing models as accounts of human behavior

We recruited 160 participants from Amazon's Mechanical Turk, who were each compensated $2.50.

To match human judgments to model performance, we asked participants to view a video of an object sliding down a ramp and colliding with another object on the ground (see Figure 4), until the 'lights were turned off' and the screen went dark. Based on this observation, participants were asked to judge whether the object on the ramp was lighter or heavier than the object on the ground. This judgment was registered using a sliding scale from "ground object much heavier" on the left to "ramp object much heavier" on the right, with "same weight" delineating the midpoint.

The experiment began with instructions to introduce participants to the task and response methodology, followed by a comprehension check and five example stimuli (identical for all participants) for familiarization. Participants then observed their set of 120 stimuli in a randomized order.

Stimuli: Stimuli were produced using the generative model as used for the training set with the following exceptions. Unlike the model training stimuli, the ground object was always fixed to be a Brick Block of consistent volume, density, and position. Participants were notified during the instructions that the ground object would not change.

We used 168 scenarios for this experiment. Of these scenarios, 120 were created such that the friction and density of the ramp object were set by the substance type using the means of the training distributions. The remaining 48 scenarios were created with density drawn from either a 'high' or 'low' density distribution. Each of these scenes with incongruent densities ('matched-incongruent') was matched to one of the scenes with congruent densities ('matched-congruent') such that the ramp object's visual texture, size, shape, and initial position were identical. The remaining 72 congruent scenarios ('normal-congruent') were used to ensure that incongruent scenes were surprising to participants.

To investigate how human mass judgments evolve over time, we made four videos from each scenario, differing in when the screen went dark: one turning black the frame before before the collision of the two objects ('pre-collision'), one changing 200ms after the collision ('post-collision'), one cutting halfway between the time when the collision occurred and when all motion would stop ('halfway'), and one that ended 200ms after both objects had come to rest ('full').

Trials were counterbalanced such that each participant only observed one video length from each of the matched trial pairs, while keeping a constant proportion of material types, shapes, and video lengths within each condition. Therefore, participants each observed 72 'normal-congruent', 24 'matched-congruent', and 24 'matched-incongruent' trials, so that the incongruent trials were only 20% of the total trials.

Empirical results: Participants were able to recover the relative weight of objects under normal conditions; across all congruent trials, average ratings were highly correlated with the true log-mass-ratio between the ramp and ground objects ($r = 0.74, t(478) = 24, p \approx 0$). Participants were also sensitive to the dynamics of the scene; excluding the pre-collision videos (where we expect no difference), the difference between hu-

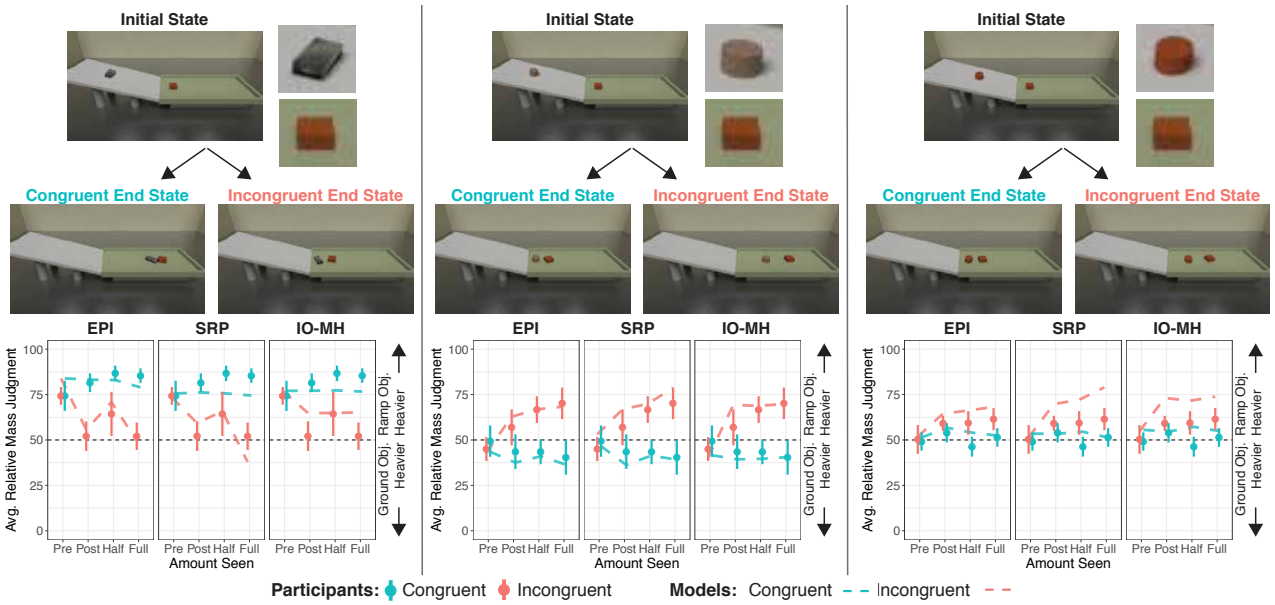

Figure 4: Three matched trials with human and model judgments. The initial image along with the end state for both the congruent and incongruent versions are displayed on top, with the ramp and ground objects magnified. Plots show average human (dots, with bars representing the 95% CI) and model (dashed line) relative weight ratings for each time point. Both the EPI and SRP models captured evolving human judgments, while the MH model typically shifted its judgments suddenly upon observing a collision.

man ratings in each matched trial was well correlated with the actual difference in log-masses of the congruent and incongruent matched objects ($r = 0.93, t(46) = 17.6, p \approx 0$). However, the difference between ratings for matched trials was not static over time (post: 15.1, half: 13.4, full: 19.7, $F(2,94) = 16.4$, $p = 7.9 * 10^{-7}$). This suggests that people incrementally integrate information about the dynamics of a scene to update their physical beliefs.

Model results: Across all scenes, participants' ratings could be predicted well by the EPI model ($r = 0.92$), the SRP model ($r = 0.93$), and IO-MH ($r = 0.88$), suggesting that all models are at least grossly approximating human behavior.

Because human weight ratings are correlated with the log-mass-ratios between the ramp and ground objects, we extrapolate "model ratings" as linear functions of the log-mass-ratio inferred by each model, with the mapping fit to the normal-congruent trials. On these trials, the SRP model explained participants' ratings best (RMSE=5.90, 95% CI=$[5.40, 6.35]$), followed closely by both the EPI model (RMSE=6.38, 95% CI=$[5.79, 6.97]$) and IO-MH model (RMSE=6.48, 95% CI=$[6.00, 7.00]$). However, when these ratings are extended to the incongruent trials, the EPI model generalized best to human ratings (RMSE=8.85, 95% CI=$[7.83, 9.63]$), followed by the SRP model (RMSE=9.81, 95% CI=$[9.05, 10.7]$) and IO-MH model (RMSE=12.3, 95% CI=$[11.0, 14.1]$).

A good model should also update it's beliefs at a similar rate as people. We compared the difference in human ratings across matched trials at each time point to the difference in model predictions of log-mass-ratios at that point. If the model is updating at the same rate as people, the linear slope between these measures should remain constant across time. There is no evidence of difference across time points for the EPI model ($F(4,138) = 0.35$, $p = 0.84$) and only limited evidence for the SRP model ($F(4,138) = 2.01$, $p = 0.096$), but clear evidence that the IO-MH model differs ($F(4,138) = 3.56$, $p = 0.010$).

Together, this suggests that the EPI model and the rational process model it approximates can capture how human weight ratings change for surprising events over time, whereas a computationally-unbounded ideal observer model cannot.

Discussion

We presented the EPI network which addresses the problem of physical reasoning with a recurrent recognition network in a generative model of physics and graphics. We found that EPI is on par or exceeds sampling based idealistic inference schemes in accuracy and efficiency. The model gave a highly accurate quantitative account of the dynamics of human mass judgments, supporting efficient inference as a mechanism underlying human physical scene understanding. Layers of the model provide a hypothesis about the neural computations underlying physical reasoning.

Acknowledgments This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216 and by a grant from Mitsubishi.

References

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

Battaglia, P. W., Pascanu, R., Lai, M., Rezende, D. J., et al. (2016). Interaction networks for learning about objects, relations and physics. In *NIPS*.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *PNAS*, *113*(34), E5072–E5081.

Sliwa, J., & Freiwald, W. A. (2017). A dedicated network for social interaction processing in the primate brain. *Science*, *356*(6339), 745–749.

Wu, J., Lu, E., Kohli, P., Freeman, W. T., & Tenenbaum, J. B. (2017). Learning to see physics via visual de-animation. In *NIPS*.