Learning physics from dynamical scenes

Tomer Ullman[1] (tomeru@mit.edu), Andreas Stuhlmüller[2] (ast@mit.edu), Noah Goodman[2] (ngoodman@stanford.edu), & Joshua Tenenbaum[1] (jbt@mit.edu)

[1]Department of Brain and Cognitive Sciences, MIT, [2]Department of Psychology, Stanford University

Abstract

Humans acquire their most basic physical concepts early in development, but continue to enrich and expand their intuitive physics throughout life as they are exposed to more and varied dynamical environments. We introduce a hierarchical Bayesian framework to explain how people can learn physical theories across multiple timescales and levels of abstraction. In contrast to previous Bayesian models of theory acquisition (Tenenbaum, Kemp, Griffiths, & Goodman, 2011), we work with more expressive *probabilistic program* representations suitable for learning the forces and properties that govern how objects interact in dynamic scenes unfolding over time. We compare our model and human learners on a challenging task of inferring novel physical laws in microworlds given short movies. People are generally able to perform this task and behave in line with model predictions. Yet they also make systematic errors suggestive of how a top-down Bayesian approach to learning might be complemented by a more bottom-up feature-based approximate inference scheme, to best explain theory learning at an algorithmic level.

Keywords: theory learning; intuitive physics; probabilistic inference; physical reasoning

Introduction

People regularly reason about the physical properties of the world around them. Glancing at a book on a table, we can rapidly tell if it is about to fall, how it will slide if pushed, tumble if it falls on a hard floor, sag if pressured, bend if bent. This ability for physical scene understanding begins to develop in infancy, and is suggested as a core component of human cognitive architecture (Spelke & Kinzler, 2007).

While some aspects of this capacity are likely innate (Baillargeon, 2002), learning also occurs at multiple levels from infancy into adulthood. Infants develop notions of containment, stability, and gravitational force over the first few months of life (Baillargeon, 2002). With exposure, young children acquire an intuitive understanding of remote controls and magnets. Most young children and adults quickly adjust to the 'unnatural physics' of many video games, and astronauts can learn to adjust to weightless environments.

How, in principle, can people learn intuitive physics from experience? How can they grasp structure at multiple levels, ranging from deep enduring laws acquired early in infancy to the wide spectrum of novel and unfamiliar dynamics that adults encounter and can adapt to? How much data are required, and how are the data brought to bear on candidate theory hypotheses? These are the questions we ask here.

We take as a starting point the computational-level view of theory learning as rational statistical inference over hierarchies of structured representations (Tenenbaum et al., 2011). Previous work in this tradition focused on relatively spare and static logical descriptions of theories and data; for example, a law of magnetism might be represented as 'if magnet(x) and magnet(y) then attract(x,y)', and the learner's data might consist of propositions such as 'attracts($object_a$, $object_b$)' (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010). Here we adopt a more expressive representational framework suitable for learning the force laws and latent properties governing how objects move and interact with each other, given observations of scenes unfolding dynamically over time. We compare the performance of an ideal Bayesian learner who can represent dynamical laws and properties with the behavior of human learners asked to infer the novel physics of various microworlds from short movies (e.g., the snapshot shown in Fig. 1). While people are generally able to perform this challenging task, they also make systematic errors which are suggestive of how they might use feature-based inference schemes to approximate ideal Bayesian inference.

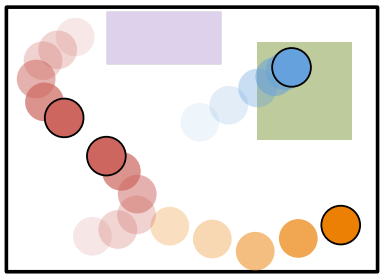

Figure 1: Illustration of the domain explored in this paper, showing the motion and interaction of four different pucks moving on a two-dimensional plane governed by latent physical properties and dynamical laws, such as mass, friction, global and pairwise forces.

Formalizing theory learning

The core of our formal treatment is a hierarchical probabilistic generative model for theories (Kemp et al., 2010; Ullman, Goodman, & Tenenbaum, 2012), specialized to the setting of intuitive physical theories (Fig.2). The hierarchy consists of several levels, with more concrete (lower-level) concepts being generated from more abstract versions in the level above, and ultimately bottoming out in data that take the form of dynamic motion stimuli. Generative knowledge at each level is represented formally using `(define ...)` statements in Church, a probabilistic programming language (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008).

Probabilistic programs are useful for representing knowledge with uncertainty (e.g. Stuhlmüller & Goodman, 2013). Fig. 2(iii) shows examples of probabilistic definition statements within our domain of intuitive physics, using Church. Fig. 2(i) shows the levels associated with these statements. The arrows from one level to the next represent how each level is sampled from the definitions and associated probability distributions of the level above it.

It is not possible to fully detail the technical aspects of the model in the space provided, and so we provide a general overview. The model is a hierarchy of levels from N (framework level) to 0 (observed data). The top-most level N

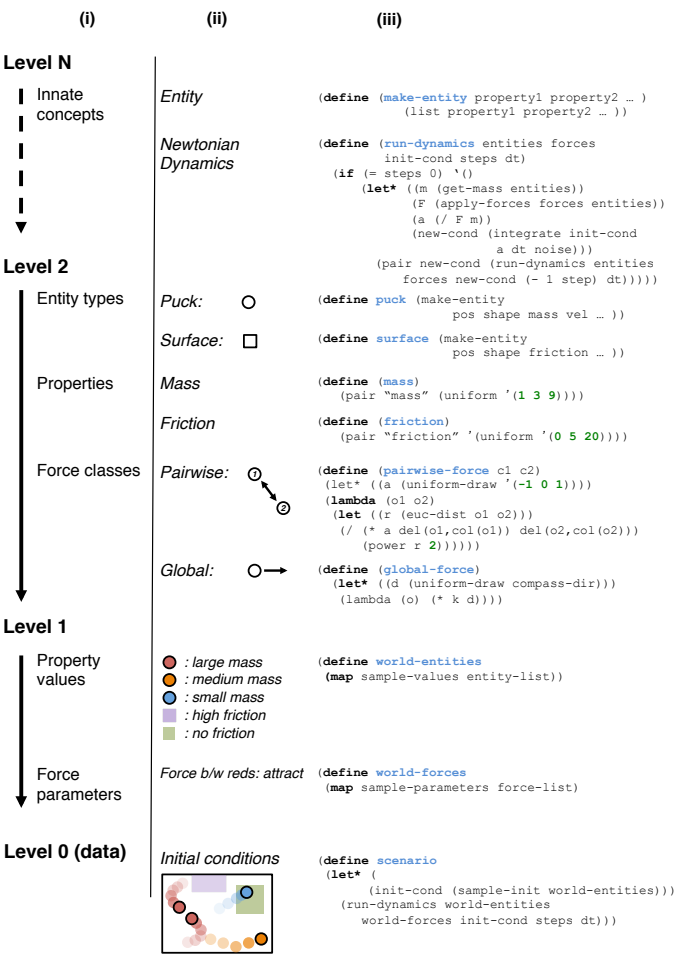

Figure 2: Formal framework for learning intuitive physics in different domains: (i) The general hierarchy going from abstract principles and assumptions to observable data. The top-most level of the hierarchy assumes a general noisy-Newtonian dynamics. (ii) Applying the principles in the left-most column to the particular domain illustrated by Fig. 1 (iii) Definition statements in Church, capturing the notions shown in the middle column with a probabilistic programming language.

represents general framework knowledge (Wellman & Gelman, 1992) and expectations about physical domains. The concepts in this level include **entities**, which are a collection of **properties**, and **forces**, which are functions of properties and govern how these properties change over time. Forces can be fields that apply uniformly in space and time, such as gravity, or can be event-based, such as the force impulses exerted between two objects during a collision. Properties are named values or distributions over values. Properties such as *location* and *shape* are privileged - it is assumed all entities have them. *Mass* is another privileged property - it is assumed all dynamic entities (those that can potentially move) have it. Dynamic entities correspond then to the common definition of matter as 'a thing with mass that occupies space'.

The framework level defines a 'Newtonian-like' dynamics, consistent with suggestions from several recent studies of intuitive physical reasoning in adults (e.g. Battaglia, Hamrick, & Tenenbaum, 2013; Sanborn, Mansinghka, & Griffiths, 2013) and infants (Téglás et al., 2011). As Sanborn et al. (2013) show, such a 'noisy-Newtonian' representation of intuitive physics can account for previous findings in dynamical perception that have supported a heuristic account of physical reasoning (Gilden & Proffitt, 1989; Todd & Warren, 1982), or direct perception models (e.g. Andersson & Runeson, 2008).

Descending from Level N to Level 0, concepts are increasingly grounded by sampling from the concepts and associated probability distributions of the level above Fig. 2(i)). Each level in the hierarchy can spawn a large number of instantiations in the level below it. Each lower level of the hierarchy details the *types* of possible entities, properties and forces in it. All members of an entity type share properties, and are governed by the same types of forces. A force type specifies the number and types of entities it acts on, and how their relevant properties change over time.

Space of learnable theories. Levels 0-2 in Fig. 2 capture the specific sub-domain of intuitive physics we study in this paper's experiments: two-dimensional discs moving over various surfaces, generating and affected by various forces, colliding elastically with each other and with barriers bounding the environment (cf Fig. 1).

Levels 0-2 represent the minimal framework needed to explain behavior in our task and we remain agnostic about more abstract background knowledge that might also be brought to bear. We give participants explicit instructions that effectively determine a single Level 2 schema for the task, which generates a large hypothesis space of candidate Level 1 theories, which they are asked to infer by using observed data at Level 0.

Level 2: The "hockey-puck" domain. This level specifies the entity types *puck* and *surface*. All entities within the type *puck* have the properties *mass*, *elasticity*, *color*, *shape*, *position*, and *velocity*. Level 2 also specifies two types of force: *Pairwise forces* cause pucks to attract or repel, following the 'inverse square distance' form of Newton's gravitation law and Coulomb's Law. *Global forces* push all pucks in a single compass direction. We assume forces of *collision* and *friction* that follow their standard forms, but they are not the subject of inference here.

Level 1: Specific theories. The hockey-puck domain can be instantiated as many different specific theories, each describing the dynamics of a different possible world in this domain. A Level 1 theory is determined by sampling particular values for all free parameters in the force types, and for all entity subtypes and their subtype properties (e.g., masses of pucks, friction coefficients of surfaces). Each of the sampled values is drawn from a probability distribution that the Level 2 theory specifies. So, Level 2 generates a prior distribution over candidate theories for possible worlds in its domain.

The domain we study here allows three types of pucks, indexed by the colors red, blue and yellow. It allows three types of surfaces (other than the default blank surface), indexed by the colors brown, green and purple. Puck mass values are 1, 3, or 9, drawn with equal probability. Surface

friction coefficients values are 0, 5 or 20, drawn with equal probability. Different pairwise forces (attraction, repulsion, or no interaction) can act between each of the different pairs of puck types, drawn with equal prior probability. Finally, a global force may push all pucks in a given direction, either $\uparrow, \downarrow, \leftarrow, \rightarrow$ or 0, drawn with equal probability. We further restrict this space by considering only Level 1 theories in which all subclasses differ in their latent properties (e.g. blue, red and yellow pucks must all have different masses). While this restriction (together with the discretization) limits the otherwise-infinite space of theories, it is still a very large space, containing 131,220 distinct theories.

Level 0: Observed data. The bottom level is a concrete scenario, specified by the precise individual entities under observation and the initial conditions of their dynamically updated properties. Each Level 1 theory can be instantiated in many different scenarios. The pucks' initial conditions were drawn from a zero-mean Gaussian distribution for positions and a Gamma distribution for velocities. Once the entities and initial conditions are set, the positions and velocities of all entities are updated according to the Level 1 theory's specific force dynamics for T time-steps, generating a path of multi-valued data points, $d_0, \ldots, d_T$. The probability of a path is simply the product of the probabilities of all the choices used to generate the scenario. Finally, the actual observed positions and velocities of all entities are assumed to be displaced from their true values by Gaussian noise.

Theory learning as Bayesian inference

The model described so far allows us to formalize different kinds of learning as inference over different levels of the hierarchy. This approach can in principle be used for reasoning about all levels of the hierarchy, including the general shape of forces and types of entities, the unobserved physical properties of entities, as well as the existence, shape and parameters of unseen dynamical rules. In this paper, we specifically consider inference over the properties of mass and friction, and the existence and direction of pairwise and global forces. We do this by inverting the generative framework to obtain the posterior over all possible theories that could have produced the observed data. We then marginalize out irrelevant aspects of the theory to obtain posterior probabilities over the dynamic quantity of of interest (Fig.3a and b).

Simulation-based approximations

The inversion of the generative model is in principle sufficient for inference over any unknown quantity of interest in it, and in our particular discretized domain we can explicitly sum over all possible theories. However, integrating over the full space of theories is generally intractable, and it is implausible that for any dynamic stimuli people perform massive inference over all possible models that could have generated it. Also, people can use more than point-wise deviation between expected paths to estimate physical parameters. For example, if people think two objects attract they might expect that over time the mean distance between the objects should shrink.

This psychological intuition suggests a principled way of approximating the full inference, following a statistical method known as Approximate Bayesian Computation (see Blum, Nunes, Prangle, & Sisson, 2013, for a review). This approach is similar to 'indirect inference' which assumes a model that can generate simulated data d' given some parameters θ, but does not try to estimate θ directly from observed data d. Rather, we first construct an auxiliary model with parameters β and an estimator $\tilde{\beta}$ that can be used to evaluate both d and d'. The indirect estimate of the parameter of interest, $\hat{\theta}$, is then the parameter that generated the simulated data whose estimator value $\tilde{\beta}(d')$ is as close as possible to the estimator value of observed data, $\tilde{\beta}(d)$.

Here we will use the following approximation: Our framework allows us to generate simulated object paths given physical parameters θ, which we then wish to estimate. We begin by drawing simulated data for all the models within the domain over all scenarios, giving us several hundred thousand paths. For every physical parameter θ we construct a set of summary statistics that can be evaluated on any given path, and act as estimators. For example, the summary statistic $avgPositionX(d)$ calculates the mean x-axis position of all objects over a given path, and can be used as an estimator for the existence of a global force along the x-axis. We evaluate these sufficient statistics for each of the parameter values over all the paths, obtaining an empirical likelihood distribution which is smoothed with gaussian kernels. The estimated likelihood of a given parameter is then the likelihood of the sufficient statistic for the observed data (see Figure 3 for an illustration of this process).

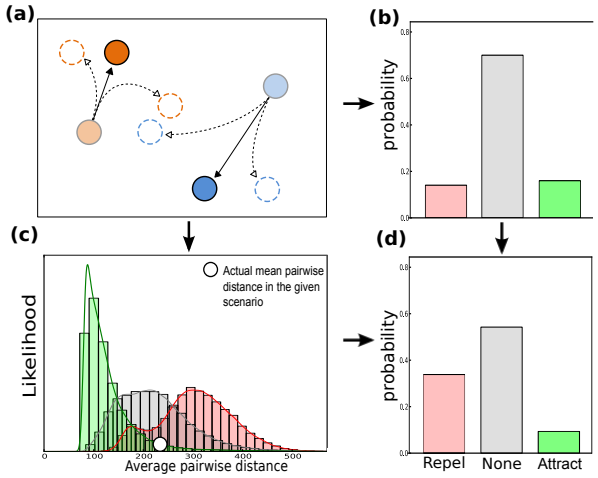

Figure 3: Approximations and the ideal observer for pairwise forces. For a given scenario (a), many alternate paths are generated and compared to the observed path, giving a log likelihood for all theories. Posterior estimates are obtained by marginalizing over the theories (b), or by comparing the summary statistics of the scenario to its empirical distribution over many simulations done offline (c). We also consider a simple linear combination of the methods (d).

Psychologically, this approximation means the following: people can imagine dynamical scenes unfolding over time, but when reasoning about a specific scene they only compute certain summary statistics. People compare the values of the

summary statistics in the specific scene to a repository which was computed by imagining many possible scenes. These repositories are built by using the same imagery capacity that lets people imagine individual scenes evolving. For example, people might watch a specific scene and compute how close some pucks are on average (the summary statistic). People then compare that value to a repository of average particle distances. People can then conclude that the particles are not attracting – because based on the repository of previous scenes if the particles were attracting they should be closer together on average.

We examine these various ways of physical reasoning, by considering people's performance on a novel dynamical task.

Experiment

Participants Three hundred participants from the US were recruited via the Amazon Mechanical Turk service, and were paid for their participation. Ten participants were excluded from analysis for failing comprehension questions.

Stimuli 60 videos were used as stimuli, each one lasting 5 seconds and depicting several pucks moving and colliding.

We constructed the stimuli in the following manner: First, we defined a set of 10 *worlds* that differ in the physical rules underlying their dynamics, as well as in the properties of the objects that appear in them. For example: in $world_1$ blue pucks have a large mass and there are no global or coupling forces, whereas in $world_5$ blue pucks are light and red pucks repel one another. A full description of the underlying physical rules of each world is available at `http://www.mit.edu/~tomeru/physics2014/underlyingRules.pdf`

Next, for each world we created 6 different *scenarios* that differ in their initial conditions (i.e. the starting location and velocity of the pucks and surfaces), as well as the particular objects used and the size of the surfaces. For example: the first scenario of $world_1$ has red, yellow and blue pucks, whereas the second scenario uses only red and yellow pucks. The initial conditions were drawn from random distributions, and in practice most of the movies started with the pucks already moving in some arbitrary way.

Given the dynamical rules of the world and initial conditions, we unfolded the scenarios over 400 steps and created a video detailing the motion of the objects over time [1].

Procedure Each participant saw 5 videos drawn from the set of 60 possible stimuli. The video-participant pairing was done according to a Latin-square design, such that approximately thirty participants saw each video. The order of the 5 videos was randomized for each participant.

Participants were informed what objects, forces and physical properties were potentially present across all the stimuli, and that objects of the same color have the same properties. Participants were instructed to think of the videos as similar to 'hockey pucks moving over a smooth white table-top'.

[1] All stimuli are available at `http://www.mit.edu/~tomeru/physics2014/stimuli/`

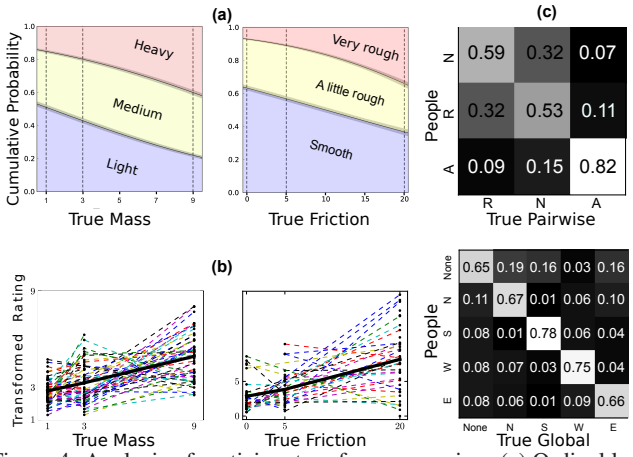

Figure 4: Analysis of participant performance using: (a) Ordinal logistic regression for mass (left) and friction (right). Shaded black areas represent uncertainty on parameter estimates, colored patches show the ordinal responses. The downward trend indicates a greater proportion of participants selecting the qualitatively correct response as the quantitative value goes up, (b) Per scenario analysis with transformed ratings for mass (left) and friction (right). Each black dot represents the average rating of 25-30 participants. The solid line shows the average response across all scenarios. Dotted lines connect mass/friction ratings in the same scenario, and so a rising line means a correct ranking. (c) Confusion matrices for pairwise forces (top) and global forces (bottom).

After the presentation of each video participants rated the entire set of possible physical properties. For example, for each puck color participants were asked how massive are [color] pucks?, with possible answers being heavy, medium or light. In some cases not all the questions were relevant (for example, a question about the mass of the blue puck when no blue pucks are present), in which case participants could answer "can't tell from video". 13 questions per video, 5 videos per per participant means 65 data points per participant. An example experiment is available at `http://www.mit.edu/~tomeru/physics-experiment-turk/physics-experiment.html`.

Participant Performance

Overview Participants correctly answered 53% of the questions on average, with a standard error of $\sim 13\%$.[2]

Participants' quantitative performance was generally good given the nature of the stimuli, and it differed depending on the particular physical property being considered.

Analysis We analyzed the results in two ways:

Aggregating over the different scenarios: We obtained the empirical distribution of responses over the possible answers across all scenarios. We collapsed across the property of color to consider 4 physical properties: Mass, friction, pairwise forces and global forces. For mass and friction properties the responses were clearly ordinal (light, medium, and heavy for mass; smooth, a little rough, and very rough for friction) and the ground truth was a continuous ratio scale, so we can fit an ordinal logistic regression to the participants,

[2] There was a small significant effect of learning over time, which we do not account for explicitly in the analysis: 52% correct on first 2 videos vs. 54% answers on last 2 videos.

shown in Fig. 4a. The figure displays the cumulative probability on the left y-axis, and the relevant response on the right y-axis. For example, on this regression the probability people will answer 'light' when the true mass is in fact light (equal to 1) is 52%. The probability they will answer 'medium' is 33% (85%-52%), and the probability that they will answer 'heavy' is the remaining 15%. This is close to the empirical values of 47%/37%/16%.

An ordinal regression cannot be used for the global and coupling forces, and so Fig. 4c shows empirical confusion matrices, detailing the percentage of people which chose each option given the ground truth.

Transforming responses per scenario For mass and friction we can assess performance in a more refined way, by considering the distribution of responses for each puck (and surface) in scenario, and transforming this distribution into a quantitative prediction. We take the expectation of the physical property relative to the empirical distribution (e.g., if 60% of participants rated a yellow puck in scenario 7 as 'heavy' and 40% rated it as 'medium', the converted rating is $0.6*9+0.4*3=6.6$), and compare with the ground truth (Fig. 4b). We next consider each property separately:

1. Mass: The downward trend in Fig. 4a, shows that participants correctly shift in the probability of answering that a mass is heavier as it becomes heavier. The linear correlation in Fig. 4b shows that despite a large degree of variance for any given mass, participants were able to overall correctly scale the masses. The ability to correctly rank and quantitatively scale multiple masses is of interest, as experiments on inferring mass from collisions usually focused on judgements of mass ratios for two masses, often requiring binary responses of 'more/less massive' (e.g. Gilden & Proffitt, 1989). **2. Friction**: Again we see a downward trend in the logistic regression, depicted in Fig. 4a. Compared with the regression for the masses, participants lean more heavily towards the lower end of the responses, perhaps because a 'null' response (no friction) is easier to make than a graded response along a continuum. The linear correlation depicted in Fig. 4b shows that participants were also able to correctly rank the roughness of the surfaces, though they could better distinguish between high- and low-friction surfaces than they were able to distinguish low- and zero-friction surfaces. To our knowledge this is the first systematic study of people's ranking of the friction properties of surfaces in the intuitive physics literature. **3. Pairwise forces**: As shown in Fig. 4c participants performed well on attraction forces, correctly detecting them on average in 82% of the cases in which they existed, while not reporting them on average in 88% of the cases in which they did not exist. As for repulsion and non-forces, their performance was above chance, although it was significantly worse than attraction. Note in particular that there is an asymmetry in the column for non-forces, indicating participants are confusing repulsion and non-existent forces, much more than they are confusing attraction and non-forces (32% vs. 15%). **4. Global forces**: As shown in Fig. 4c participants performed relatively well on detecting global forces, identifying the correct global force 70% of the time on average. Note that generally any force is more likely to be confused with the null-force case than it is with any specific force.

Comparison to different models

For the *Ideal Observer* model (IO), we get predictions in the following way: For each scenario, we fix the observed initial conditions and simulate the resulting paths under all parameter hypotheses. We then score each model by assessing the deviation of its simulated path from the observed path. Finally, for each parameter of interest we marginalize over the other parameters, to obtain a log-likelihood score for the relevant parameter (see Fig. 3a and b). For the *Simulation Based Approximation* model (SBA), we get predictions by following the procedure detailed at the end of the formal modeling section. We also consider a combination of these two approaches, linearly summing weighted likelihoods from both approaches for any given physical parameters. These various approaches are illustrated for a particular example in (see Fig. 3). All models are fit to participant data by uniformly scaling their predictions using a noise parameter β, optimizing for root-mean-square-error (RMSE).

We begin our comparison by collapsing across scenarios to compare with the logistic regressions and confusion matrices shown in Fig. 4. For **mass** inference, the SBA model outperforms the IO model and is quite close to people's performance. The combined IO&SBA model places its entire weight on the SBA model. For **friction**, the SBA model outperforms the IO model, although the combined IO&SBA outperforms both (Fig.5a). For the confusion matrices we measured fit using RMSE. For **pairwise** forces, people showed a particular asymmetry when they incorrectly judge a null-force, mistaking it more often for a repulsive force than an attractive one (Fig. 4b). We can understand this difference intuitively – attraction pulls bodies together which provides even further evidence for attraction over time, while repulsion pushes bodies apart and becomes weaker, providing less evidence for itself over time. Such an asymmetry plays out over the entire scene, and does not come naturally out of the IO model, which sums up error across local deviations. By contrast, a summary statistic measuring the average pairwise distance does replicate this asymmetry. The combined IO&SBA is the closest to that of people in terms of RMSE (Fig. 5b). For **global** forces, people were confused between any given force and the absence of force, relative to any other force. Both the IO and SBA models replicate this finding, although the IO model is closer to people. Also, the SBA model is quite bad at detecting the absence of global forces, perhaps because there is no simple feature to account for a null-force. Again, the combination IO&SBA produces a confusion matrix which is closest to that of people (Fig. 5c).

We also considered the correlation between people and the models for each scenario, for each object and property, without refitting the noise parameter. We found that for most physical properties the combined model's results were equal

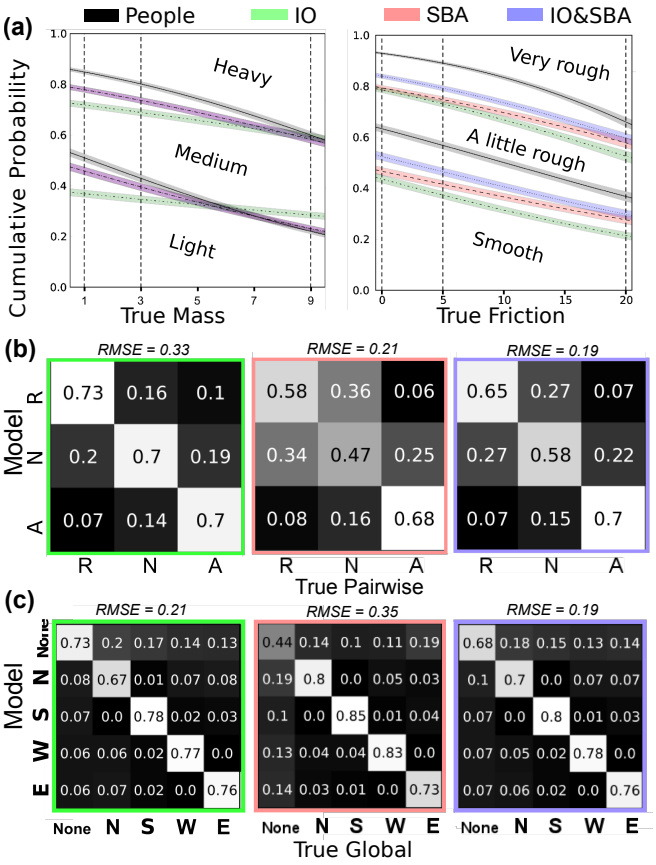

Figure 5: Comparison of model performance for properties (a) friction and mass (b) pairwise forces and (c) global forces.

to or slightly better than the individual models. For the pairwise forces there was a significant gain, with the correlation for IO&SBA being 0.81, while the SBA model and IO model had correlations of 0.75 and 0.56 respectively.

General discussion

We found that a hybrid between top-down Bayesian learning and bottom-up feature-based heuristic inference emerged as the best empirical fit to participants' behavior in learning physical laws from dynamic scenes. This general approach makes good engineering sense: It can transcend inherent limitations of each component method and serve as the basis for more robust real-world learning. The ideal Bayesian observer uses evidence in an optimal way, but it is computationally intractable. The feature-based statistics are a useful heuristic in many cases, but are unable to account for the inferences people make when the initial conditions of a scenario deviate from the norm. In our setup, feature-based statistics do not replace the knowledge of a generative model, since they themselves require the simulations of a generative model to be computed. We considered a simple way of linearly combining the top-down and bottom-up models. While this approach performed reasonably, it does not get around the need to search a large space of theories for the ideal observer. A more psychologically plausible mechanism might include using the summary statistics of a given scenario to pick out a small space of 'reasonable' theories and then use Bayesian inference on this smaller space.

There are many questions that are still open when considering the challenge of inferring physical dynamics from perceptual scenes. For example, to what extent are the computational processes underlying intuitive physics shared between adults and children? While some physical knowledge develops over time, it is possible that a basic understanding of entities, forces and dynamics, is innate. Our experiments focused on adults, but one advantage of our novel stimuli is that they can be easily adapted to experiments with young children or infants, using simple responses or violation-of-expectation to indicate what they learn from brief exposures.

The combination of hierarchical Bayesian learning, an expressive representation for dynamical theories in terms of probabilistic programs, and psychologically plausible approximate inference schemes offers a powerful paradigm for modeling the content and acquisition of a broad swath of human intuitive physics.

Acknowledgments This material is based on work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216, and ONR grant N00014-13-1-0333.

References

Andersson, I. E., & Runeson, S. (2008). Realism of confidence, modes of apprehension, and variable-use in visual discrimination of relative mass. *Ecological psychology*, *20*(1), 1–31.

Baillargeon, R. (2002). The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell handbook of childhood cognitive development*, 47–83.

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.

Blum, M., Nunes, M., Prangle, D., & Sisson, S. (2013). A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science*, *28*(2), 189–208.

Gilden, D. L., & Proffitt, D. R. (1989). Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 372–383.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: a language for generative models. *Uncertainty in Artificial Intelligence*.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165 – 196.

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

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Stuhlmüller, A., & Goodman, N. D. (2013). Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*, 1054-1059.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Todd, J. T., & Warren, W. H. (1982). Visual perception of relative mass in dynamic events. *Perception*, *11*(3), 325–335.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455 - 480.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual review of psychology*, *43*(1), 337–375.