Cognitive Psychology 104 (2018) 57–82

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Learning physical parameters from dynamic scenes

Tomer D. Ullman [a,*], Andreas Stuhlmüller [a], Noah D. Goodman [b], Joshua B. Tenenbaum [a]

[a] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, USA
[b] Department of Psychology, Stanford University, Stanford, USA

ARTICLE INFO

Article history:
Accepted 26 May 2017
Available online 11 April 2018

Keywords:
Learning
Intuitive physics
Probabilistic inference
Physical reasoning
Intuitive theory

ABSTRACT

Humans acquire their most basic physical concepts early in development, and 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 parameters at multiple levels. 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 to human learners on a challenging task of estimating multiple physical parameters in novel microworlds given short movies. This task requires people to reason simultaneously about multiple interacting physical laws and properties. People are generally able to learn in this setting and are consistent in their judgments. Yet they also make systematic errors indicative of the approximations people might make in solving this computationally demanding problem with limited computational resources. We propose two approximations that complement the top-down Bayesian approach. One approximation model relies on a more bottom-up feature-based inference scheme. The second approximation combines the strengths of the bottom-up and top-down approaches, by taking the feature-based inference as its point of departure for a search in physical-parameter space.

1. Introduction

Reasoning about the physical properties of the world around us is a ubiquitous feature of human mental life. Not a moment passes when we are not, at least at some implicit level, making physical inferences and predictions. Glancing at a book on a table, we can rapidly tell if it is about to fall, or how it will slide if pushed, tumble if it falls on a hard floor, sag if pressured, bend if bent. The capacity for physical scene understanding begins to develop early in infancy, and has been suggested as a core component of human cognitive architecture (Spelke & Kinzler, 2007).

While some parts of this capacity are likely innate (Baillargeon, 2008), learning also occurs at multiple levels from infancy into adulthood. Infants develop notions of containment, support, stability, and gravitational force over the first few months of life (Baillargeon, 2002; Needham & Baillargeon, 1993), as well as differentiating between liquid substances and solid objects (Hespos, Ferry, & Rips, 2009; Rips & Hespos, 2015). Young children build an intuitive understanding of remote controls, magnets, touch screens, and other physical devices that did not exist over most of our evolutionary history. Astronauts

* Corresponding author.
E-mail address: tomeru@mit.edu (T.D. Ullman).

https://doi.org/10.1016/j.cogpsych.2017.05.006
0010-0285/

and undersea divers learn to adapt to weightless or partially weightless environments (McIntyre, Zago, Berthoz, & Lacquaniti, 2001), and video-game players can adjust to a wide range of game worlds with physical laws differing from our everyday natural experience.

Not only can we learn or adapt our intuitive physics, but we also seem to do so from remarkably limited and impoverished data. While extensive experience may be necessary to achieve expertise and fluency, only a few exposures are sufficient to grasp the basics of how a touch screen device works, or to recognize the main ways in which a zero-gravity environment differs from a terrestrial one. While active intervention and experimentation can be valuable in discovering hidden causal structure, they are often not necessary; observation alone is sufficient to infer how these and many aspects of physics operate. People can also gain an intuitive appreciation of physical phenomena which they can only observe or interact with indirectly, such as the dynamics of weather fronts, ocean waves, volcanoes or geysers.

Several questions naturally follow. How, in principle, can people learn aspects of intuitive physics from experience? What is the form of the knowledge that they learn? 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, and what kind of data are required for learning different aspects of physics, and how are the data brought to bear on candidate hypotheses? In this paper we present a theoretical framework that aims to begin to answer these questions in computational terms, and a large-scale behavioral experiment that tests the lower-levels of the framework, as an account of how people estimate basic aspects of physical dynamics from brief moving scenes.

Our modeling framework takes as a starting point the computational-level view of theory learning as rational statistical inference over hierarchies of structured representations (Gopnik & Wellman, 2012; Tenenbaum, Kemp, Griffiths, & Goodman, 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. Our representation includes logical machinery to express abstract properties and laws, but also numerical and vector resources needed to express the observable trajectories of objects in motion, and the underlying force dynamics causally responsible for those motions. We can express all of this knowledge in terms of a *probabilistic program* in a language such as Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008; Goodman, Tenenbaum, & Gerstenberg, 2015).

In addition to extending the framework of learning as inference over generative programs, this work follows a growing body of work that represents intuitive physical notions using probabilistic programs and data structures similar to those used in game "physics engines": computer software that efficiently simulates approximate Newtonian interactions between large numbers of objects in real time for the purposes of interactive video games (Battaglia, Hamrick, & Tenenbaum, 2013; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). This framework is different from previously proposed feature-based or heuristic-based decision rules for intuitive physics (see e.g. Gilden & Proffitt, 1989; Todd & Warren, 1982), as well as more recent models based on Bayesian networks (although it is certainly more closely related to the latter). A contrastive illustration of the kind of dynamic scenes we study – and the accompanying representation – is shown in Fig. 1. Previous research on learning physics from dynamical scenes has tended to focus on the inference of object properties under known force laws, and typically on only the simplest cases, such as inferring the relative mass of two objects in motion from observing a single collision between them, usually with one object starting at rest (see for example Andersson & Runeson, 2008; Gilden &

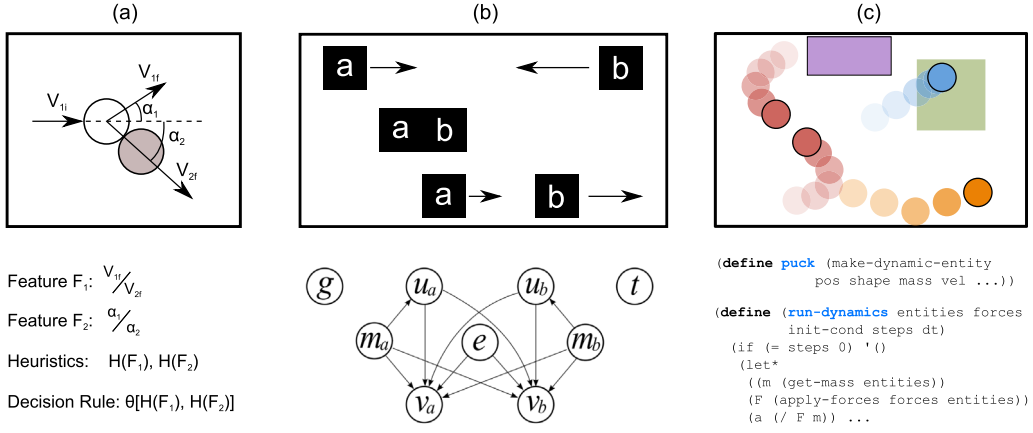

Fig. 1. Illustration of several previous physics domains and models, and the current domain and model. (a) Heuristic-based decision model for judging mass from two-dimensional collisions with one object at rest (adapted from Gilden and Proffitt, 1989) (b) Causal Bayes net for judging mass from one-dimensional collisions with both bodies in motion (adapted from Sanborn et al., 2013) (c) Part of a generative-program based model for inferring the properties of multiple pucks moving and interacting on a two-dimensional plane, governed by latent physical properties and dynamical laws, such as mass, friction, global forces and pairwise forces.

Proffitt, 1989; Runeson, Juslin, & Olsson, 2000; Todd & Warren, 1982 and the top row of Fig. 1a and b). People's inferences in these studies have traditionally been modeled as combinations of decision-rules based on different heuristics and features, or more recently, Sanborn, Mansinghka, and Griffiths (2013) have proposed a Bayesian network framework for capturing similar inferences. In both heuristic and Bayesian network models, however, the accounts proposed apply directly to just one inferential scenario, such as inferring a single property (relative mass) from a single collision between two objects (bottom row of Fig. 1a and b). Neither the experiments nor the models have attempted to elucidate general mechanisms that people might use to infer multiple properties of multiple objects from complex dynamic scenes that could involve many interactions over time. That is our goal here.

Specifically, we consider a much more complex, challenging learning task that unfolds in scenarios such as that illustrated in the top row of Fig. 1c. Imagine this as something like an air hockey table viewed from above. There are four disk-shaped "pucks" moving in a two-dimensional rectangular environment under the influence of various causal laws and causally relevant properties. In a physical domain the causal laws are *force* laws, and these forces may be either local and pairwise (analogous to the way two magnetic objects typically interact) or global (analogous to the way gravity operates in our typical environment). The properties are physical properties that determine how forces act on objects, and may include both object-based and surface-based properties, analogous to inertial mass and friction respectively. A child or adult looking at such a display might come to a conclusion such as 'red pucks attract one another' or 'green patches slow down objects.' With the right configuration different physical properties begin to interact, such that an object might be seen as heavy, but in the presence of a patch that slowed it down its 'heaviness' might be explained away by the roughness of the patch.

Such dynamical displays are still far simpler than the natural scenes people see from early in development, but they are much richer than what has been studied in previous experiments on learning intuitive physics, and learning in intuitive causal theories more generally.[1] Accordingly, we consider a richer representation that can accommodate multiple objects and entities, as part of a generative probabilistic program (illustrated in the bottom row of Fig. 1c and detailed in the next section). Our framework describes several levels of a general hierarchy for estimating and learning physical parameters at different levels, but the contribution of the higher levels remains as a theoretical alternative to the frameworks discussed above. The dynamic displays considered here will test only the lower levels of the general hierarchy, much in the same way that exploring how people infer relative mass from single one-dimensional collisions are a test for a more general proposal regarding the use of features and heuristics in estimating physical parameters.

Some research on causal learning more generally has looked at the joint inference of causal laws and object attributes, but only in the presence of simple discrete events rather than a rich dynamical scene (Gopnik & Schulz, 2004; Gopnik & Sobel, 2000; Griffiths, Baraff, & Tenenbaum, 2004). For example, from observing that a "blicket-detector" lights up when objects A or B are placed on it alone or together, but does not light up when objects C or D are placed on it alone or in combination with A or B, people may infer that only objects A and B are blickets, and that the blicket detector only lights up when all the objects on it are blickets (Lucas, Bridgers, Griffiths, & Gopnik, 2014). It is not clear that studying how people learn from a small number of isolated discrete events presented deliberately and pedagogically generalizes to how they learn physics properties in the real world, where configurations of objects move continuously in space and time, and interact in complex ways that are hard to demarcate or discretize.

In this sense our experiments are intended to capture more of how we learn and estimate the physics of the real world. Participants observe multiple objects in motion over a period of five seconds, during which the objects typically collide multiple times with each other as well as with stationary obstacles, pass over surfaces with different frictional properties, and move with widely varying velocities and accelerations. We compare the performance of human learners in these scenarios with the performance of an ideal Bayesian learner who can represent precisely the dynamical laws and properties at work in these stimuli. While people are generally able to perform this challenging task in ways broadly consistent with an ideal observer model, they also make systematic errors which are suggestive of how they might use feature-based inference schemes to approximate ideal Bayesian inference. Hence we also compare people's performance to a more feature-based model. Finally, we propose a rational approximation model that combines the strengths of the ideal and feature-based inferences into a more psychologically plausible account that is based on both heuristics and an implicit understanding of Newtonian-like mechanics.

2. Formalizing physics learning

The core of our formal treatment is a hierarchical probabilistic generative model for theories (Goodman, Ullman, & Tenenbaum, 2011; Kemp et al., 2010; Ullman, Goodman, & Tenenbaum, 2012), specialized to the domain 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. While in this work we only test the framework on the lower and more concrete levels, we detail the higher level as a theoretical proposal. This theoretical proposal conveys a commitment to a knowledge representation that is in contrast to the heuristic-based decision rules and a Bayes-net formalisms for intuitive physics mentioned in the introduction. The

[1] The tradition of 'qualitative physics' has considered more complex inferences, but has focused more on qualitative scene descriptions and sketches (Forbus, 1988).

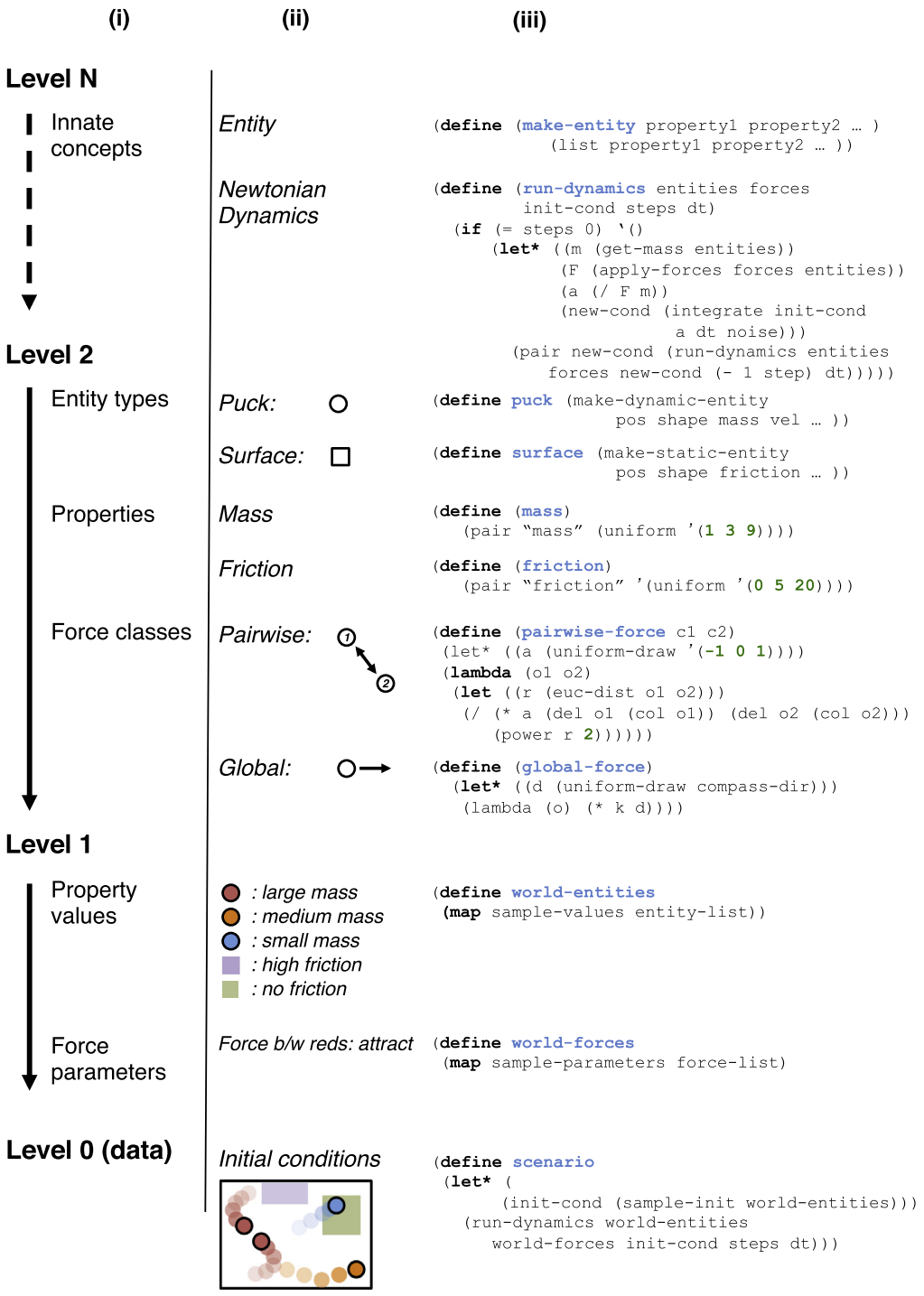

Fig. 2. Formal framework for learning intuitive physics in different domains: (i) The general hierarchy going from abstract principles and assumptions to observable data. The top-most level of the hierarchy assumes a general noisy-Newtonian dynamics. (ii) Applying the principles in the left-most column 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.

kind of knowledge representation specified below, in the form of probabilistic programs over dynamics, is proposed as a way for reasoning and learning about multiple interacting aspects of physics from complex ongoing interactions – not just learning about single physical parameters from single, isolated interactions as explored previously.

In our framework, generative knowledge at each level is represented formally using `(define ...)` statements in Church, a stochastic programming language (Goodman et al., 2008). The `(define x v)` statement binds the value v to the variable x, much as the statement `a = 3` binds the value 3 to the variable `a` in many programming languages. In probabilistic programming, however, we often bind variables with values that come from probability distributions, and thus on each run of the program the variable might have a different value. For example, `(define dice (uniform-draw l 6))` stochastically assigns a value between 1 and 6 to the variable `dice`. Whenever the program is run, a different value is sampled and assigned to `dice`, drawing from the uniform distribution.

Probabilistic programs are useful for representing knowledge with uncertainty (see for example Goodman et al., 2008; Goodman & Stuhlmüller, 2013; 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, and the arrows from one level to the next show that each level is sampled from the definitions and associated probability distributions of the level above it. The definition statements provide a formalization of the main parts of the model. The full forward generative model implementing levels 2 to 0 is available at http://forestdb.org/models/learning-physics.html.

In the text below we explain these ideas further, using informal English descriptions whenever possible, but see Goodman et al. (2008) for a more formal treatment of the programming language Church, and probabilistic programming in general. We emphasize that we are not committed to Church in particular as the proposed underlying psychological representation, but rather to the notion of probabilistic programs as cognitive representations.

Framework level. The top-most level N 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 or the forces of kinetic friction between two objects moving over each other.

Properties are named values or distributions over values. While different entities can have any number of properties, a small set of properties are "privileged": it is assumed all entities have them. In our setup, the properties *location* and *shape* are privileged in this sense.

Entities are further divided into 'static' and 'dynamic.' Dynamic entities are those that can potentially move, and all dynamic entities have the privileged property *mass*. Dynamic entities correspond then to the common sense definition of matter as 'a thing with mass that occupies space.[2]'

The framework level assumes a 'Newtonian-like' dynamics, where acceleration is proportional to the sum of the forces acting on an object's position relative to the object's mass. This is consistent with suggestions from several recent studies of intuitive physical reasoning in adults (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Sanborn et al., 2013; Smith & Vul, 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, 1994; Todd & Warren, 1982), or direct perception models (Andersson & Runeson, 2008; Runeson et al., 2000). This basic assumption regarding dynamics raises the question of whether and how Newtonian-like dynamics are themselves learned. We do not attempt to solve this question here, assuming instead that this level is either innately established or learned very early (Spelke & Kinzler, 2007), through mechanisms outside the scope of this paper.

Descending the hierarchy. 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 contains more specific entities, properties, and forces than the level above it. An example of moving from Level N to Level N-1 would be grounding the general concepts of entities and forces as more specifically 2-dimensional masses acting under collisions. An alternative would ground the same general entities and forces as 3-dimensional masses acting under conservation forces. This grounding can proceed through an indeterminate number of levels, until it ultimately grounds out in observable data (Level 0).

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 being 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 help 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*

[2] The static/dynamic distinction is motivated by similar atomic choices in most computer physics engines that are used for approximate dynamic simulations, engines that were suggested as models of human intuitive physics (e.g. Battaglia et al., 2013). In these physics engines the static/dynamic divide allows computational speed-up and memory conservation, since many forces and properties don't have to be calculated or updated for static entities. It is an interesting possibility that the same kind of short-cuts developed by engineers trying to quickly simulate physical models might also represent a cognitive distinction (Ullman et al., 2017). Similar notions have been proposed in cognitive development in the separation of 'objects' from more stable 'landscapes' (Lee & Spelke, 2010).

forces cause pucks to attract or repel, following the inverse square 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.[3]

Level 0: Observed data. The bottom level of our hierarchical model (Fig. 2) 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, and filtered to remove cases in which the pucks began in overlap. 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 then 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.

The framework grounds out at the level of observed data, and it is assumed that all of the models discussed below have reliable access to a basic input representation that includes object position, identity, velocity, size, shape, and collision-detection. Thus, the models below are not 'end-to-end' in the sense of learning from a pixel-based representation. Several recent machine-learning based approaches to physics-learning have also found that an initial underlying representation of object position, velocity, and interaction (Battaglia, Pascanu, Lai, Jimenez Rezende, & Koray, 2016; Chang, Ullman, Torralba, & Tenenbaum, 2017) is better suited for learning physical relations and properties than a pixel-based representation.

2.1. Learning physical parameters as ideal Bayesian inference: the IO model

Having specified our overall generative model, and the particular version of it underlying our "hockey puck" domain, we now turn to the question of learning. 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 (although it cannot in present form be used to reason about the assumption of Newtonian-like dynamics). In this paper, we constrain the inference to the specific levels considered in our experiment. Given observations, an ideal learner can invert the generative framework to obtain the posterior over all possible theories that could have produced the observed data. We can then marginalize out nuisance parameters (other irrelevant aspects of the theory) to obtain posterior probabilities over the dynamic quantity of interest. In the following sections we refer to this learning model as the Ideal Observer (IO).

Inference at multiple levels includes both continuous parameter estimation (e.g. the strength of an inverse-square attractive force or the exact mass value of an object) and more discrete notions of structure and form (e.g. the very existence and shape of an attractive force, the fact that an object has a certain property). This parallels a distinction between two modes of learning. The distinction appears in AI research as well as cognitive development, where it is referred to as parameter setting vs. conceptual change (Carey, 2004). In general, inferring structure and form (or conceptual change) is seen as harder than parameter estimation.

Learning at different levels could unfold over different spans of time depending on the size and shape of the learning space, as well as on background knowledge and the available evidence. Estimating the mass of an object from a well-known class in a familiar setting could take adults under a second, while understanding that there is a general gravitational force pulling things downwards, given little initial data, might take infants several months to grasp (Kim & Spelke, 1992). In this paper we consider learning at a mid-point between these two extremes, between inferring basic physical knowledge and estimating familiar parameters in a familiar environment. Our experiments involve joint estimation of multiple parameters and rudimentary structure learning in the form of discrete structural relations (pairwise and global forces), but not the more

[3] More precisely, the cross product $N(mass)! \times N(friction\ coefficients)! \times N(direction) \times N(pairwise\ combination)^{N(force\ constant)} = 131,220$. Selecting the right theory in this space is equivalent to correctly choosing 17 independent binary choices.

abstract conceptual change that could take longer and require more evidence. The basic structure of noisy Newtonian mechanics and the structure of the task is assumed present, and so we examine learning at Level 1 – the sort of learning that could happen over several seconds in a novel setting.

Bayesian inversion of the generative model is in principle sufficient for inference over any unknown quantity of interest at this level: As new data comes in, probabilities are shifted among hypotheses towards those that best balance between the likelihood of the new data, and the prior distribution over the hypothesis space. Similar to other ideal-observer models, this model describes learning at the computational level, in the Marr-Poggio sense of levels of analysis (Marr & Poggio, 1976). This account is not meant to capture the algorithmic process that people go through when learning new hypotheses. Implementing this computational model directly, by enumerating all hypotheses in the space of possible theories and computing their exact probabilities at each stage of inference, would be extremely resource demanding. To get a sense for how large this space is, in our experiments, each scenario contained four pucks and two surfaces. This restricts the number of hypotheses an ideal observer needs to consider to a maximum of 14,580 for any one scenario, out of the larger in-principle space of 131,220 theories that could describe any of our worlds. It is unlikely that people have parallel access to anything like this entire hypothesis space, or that they could explicitly consider most of these theories in a serial fashion, given the short timeframe they have for judgment.

In the next two subsections we consider two simpler, psychologically plausible algorithmic-level approximations to these ideal observer inferences. Both can be seen as "rational process models" (Griffiths, Vul, & Sanborn, 2012): efficient algorithms that can be justified on rational engineering grounds as an approximation to the rational inference framework of the Bayesian ideal observer.

2.2. Simulation based approximations and summary statistics: The SSS model

Our first psychologically plausible approximation to ideal inference uses both summary statistics, and the ability to imagine new dynamic scenes. We are motivated by the intuition that people might rely on simpler, more global forms of information to estimate physical parameters, beyond the data-rich details of local space-time paths that objects follow. For example, if people think two objects attract, they might reasonably expect that over several seconds the objects will tend to get closer, in addition to predicting at a more fine-grained level that at each time step, the distance between the objects should decrease by a specific amount.

A principled approximation to ideal Bayesian inference based on this intuition can be formulated in the spirit of the statistical method known as Approximate Bayesian Computation (see Blum, Nunes, Prangle, & Sisson, 2013 for a review). The particular technical details of the approach are not crucial for understanding the rest of the approximation model details, but put formally this approach is similar to 'indirect inference' (Gourieroux, Monfort, & Renault, 1993), 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 evaluated on both d and d'. The indirect estimate of the parameters of interest, $\hat{\theta}$, is then the parameters that generated the simulated data whose estimator value $\tilde{\beta}(d')$ is as close as possible to the estimator value of observed data, $\tilde{\beta}(d)$ (for additional technical details see for example Gourieroux et al., 1993). In less formal terms, this inference scheme works by assuming we have a model that is easy to generate new data from, once we have fixed the parameters of interest. However, evaluating the parameters of interest from observed data might be hard. Instead of directly evaluating these parameters, one generates simulated data from the model, and picks the parameters that generated simulated data that is most similar to the observed data, where similarity is measured as the distance between summary statistics of the observed and simulated data.

Here we will use the following approximation: Our framework can be used to generate simulated object paths given hidden physical parameters θ, which are also our inference target. For every physical parameter θ_i we construct a set of easy-to-calculate 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 summary statistics by sampling simulated path-data for all models within our domain, giving us several hundred-thousand paths. We then calculate the summary statistic over all the paths, obtaining an empirical likelihood distribution (conditioned on the parameter values) which is smoothed with Gaussian kernels.

When a new scene is observed, the estimated probability of a physical parameter of interest is the normalized likelihood of the summary statistic calculated on the observed data (see Fig. 3a and b for an illustration of this process). In the following sections we refer to this learning model as the Simulation and Summary Statistics model, or SSS.

Psychologically, this approximation corresponds to the following: people can imagine dynamical scenes unfolding over time, but when reasoning about a specific scene they do not imagine how the same scene could have unfolded under all the different unknown variables they are reasoning about. Instead, they compute some simple summary statistics of the specific scene, such as how close some pucks are on average. People then compare the value of these summary statistics to other simulations over possible scenes. This approximation relies on imagery, imagination, and simulation, rather than obtaining direct experience of hundreds of thousands of different scenarios and building different features to use as classifiers of theories. In this paper, we considered the limit distributions of the imaginative process by taking hundreds

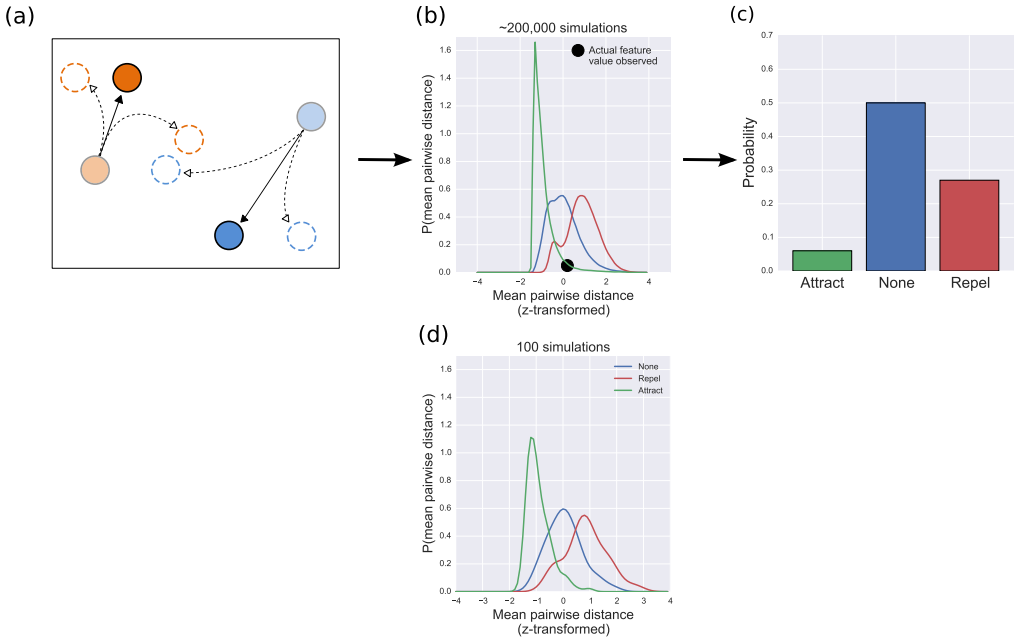

Fig. 3. Summary statistics approximation to the ideal observer for pairwise forces. For a given scenario (a) a summary statistic is computed (in this case, average pairwise distance), and then compared to (b) the empirical distribution of that summary statistic under different settings of a physical property (in this case, different settings of the pairwise force), which provides (c) the posterior distribution for a property of interest. While we considered the limit case of the empirical distribution by drawing many simulations, a smaller number of simulations can approximate this limit case. As shown in (d), the distribution obtained after 100 simulations is not far from the distribution obtained after more than 200,000 simulations.

of thousands of simulated paths. However, taking several hundred simulations (an operation on the order of seconds or less with modern hardware) can provide a reasonable approximation to these limit distributions, as illustrated in Fig. 3.

Our set of summary statistics included:

1. Average position along the x-axis
2. Total change along the x-axis
3. Average position along the y-axis
4. Total change along the y-axis
5. Average pairwise distance between particles
6. Total change in pairwise distance
7. Velocity loss while on surfaces
8. Rest time on surfaces
9. Average velocity
10. Pre- and post-collision velocity ratio
11. Change in angle following collision

The last three statistics were chosen based on heuristic models for mass judgment (Gilden & Proffitt, 1989, 1994). In the results section, we analyze these features with respect to participant performance separately of a specific model.

These summary statistics are meant to capture a large amount of possible perceptual data in the stimuli, but they are not meant to be exhaustive. We take up the question of additional possible summary statistics again in the general discussion.

While indirect inference and approximation techniques are useful, they have certain limitations, such as being insensitive to the particular conditions in outlying scenarios. That is, for any given summary statistic it is easy to construct a simple scenario that is unlikely under the statistic's likelihood, and yet people will be able to reason about without difficulty. An interesting possibility is to combine the strengths of the ideal observer model described in the previous section together with summary statistics. We suggest one such possibility below.

Finally, we stress that this approximation technique is not an alternative to the idea of inference through simulation, but rather a potentially valuable supplement to it. The simulation-based approach and related approximation is in contrast to a different possible way of approximately scoring theories, which is to learn through experience associations between theories and many features. This would require a great deal of experience, which people are unlikely to come by for the synthetic scenarios considered here, for example. This contrast is similar to the debates about top-down vs. bottom-up techniques in object perception, between those who stress a more top-down approach that relies an on actual 3D object model, and those who stress bottom-up perceptual cues calculated from still images and used for classification.

2.3. Smart initialization and short search: The START model

The two models proposed so far, the ideal observer model (IO) and the summary-statistics approximation (SSS) each have complementary advantages and disadvantages. The ideal observer is accurate but algorithmically intractable, if the learner truly evaluates all possible parameterized physical theories to compute their Bayesian score. The summary statistics are fast to compute, but are not guaranteed to learn physical parameters accurately; the approximate samples generated by SSS are not guaranteed to be an unbiased sample from the Bayesian posterior. We now consider a third model that combines their strengths. We call this the START model, for "STochastic search with smART initializations." The notion of theory learning as a stochastic search has been proposed as one way of connecting Bayesian models of cognitive development with the dynamics of children's learning (Bonawitz, Denison, Griffiths, & Gopnik, 2014; Ullman et al., 2012), but it also offers a promising way to think about learning in our dynamic physics tasks.

The START model is a psychologically plausible approximation to the ideal observer's inference, based on sampling hypotheses in a way that approaches the true Bayesian posterior as more samples are drawn. Such sampling procedures are often invoked as the logic behind algorithmic-level rational process accounts of Bayesian inference, from decision-making to perception (Gershman, Horvitz, & Tenenbaum, 2015; Gershman, Vul, & Tenenbaum, 2012; Griffiths et al., 2012; Vul, Goodman, Griffiths, & Tenenbaum, 2014). In particular, we consider a sampling technique based on Markov Chain Monte Carlo (MCMC). Informally, MCMC captures the intuition that learners consider only one hypothesis at a time, but sequentially propose changes to their current best theory, searching locally through the same space of hypotheses that an ideal observer might evaluate in parallel. The proposed theory changes are stochastic, and they are accepted or rejected stochastically according to whether they improve or decrease the posterior probability of the theory (according to the ideal observer, as described above in the section on Bayesian learning). Even if the search is initialized to a randomly chosen hypothesis, it is guaranteed to eventually converge to high posterior probability theories. It will approximate the ideal observer model's inferences if the learner is willing to search long enough. However, that search might have to be extremely long: tens of thousands or even millions of stochastic proposals (MCMC iterations) might be required. If its samples are limited, then the learner may deviate systematically from the ideal observer. See Fig. 4a–c for an illustration of this process.

In the START model, we use stochastic search with a smart initialization, which allows reasonably accurate inferences to be made by a bounded learner after only a short search, consisting of just several hundred MCMC samples. How can the learner guess an initial point for their search through theory space that is reasonably close to the best theory, before explicitly evaluating any candidates, and without access to the entire hypothesis space? This is where the summary statistics from the SSS model provide a valuable guide.

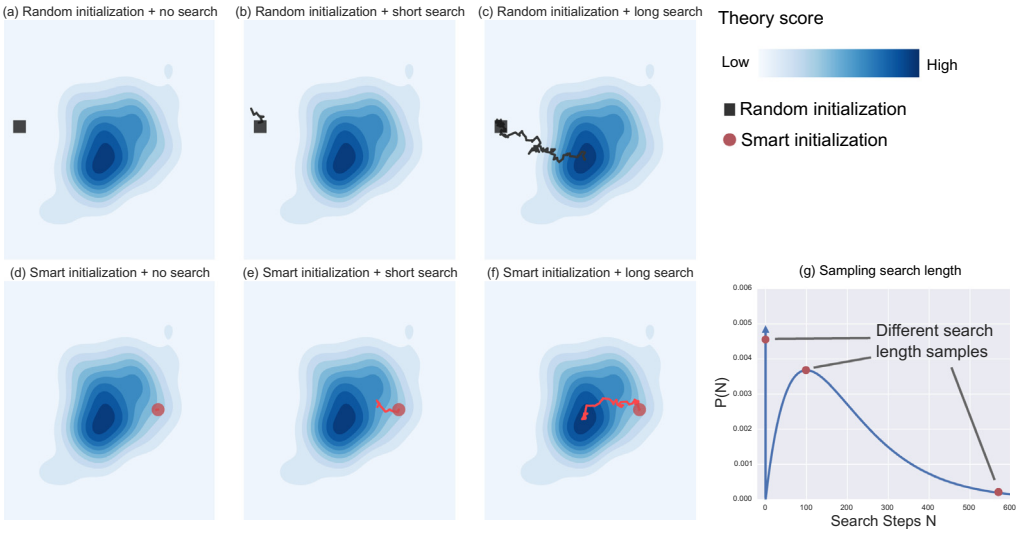

Fig. 4. Illustration of different processes for searching a space of theories. A space of possible (physical) theories is illustrated as a two-dimensional grid, where each point represents a particular theory, and color lightness indicates the score of that theory relative to the data. In a standard algorithmic approximation to an ideal observer model (a–c), the search algorithm is randomly initialized with a specific theory, and then proceeds to search the space by proposing new theories, and comparing their score to the current theory. Using a small number of search steps (b) is unlikely to reach probable theories, but in the limit of many search steps (c) this process approximates the ideal observer. Rather than a random initialization, a learner can use a 'smart initialization', informed by the summary statistics (d–f). In the START model, the number of search steps N to carry out following this smart initialization is sampled from the distribution shown in (g), and described in the main text. The distribution also shows three example samples corresponding to different times shown in (d–f). A bounded learner with limited resources can use smart initialization and either remain with the resulting theory (d) or carry out a short search (e), as a fast rational approximation to the full model. If this learner carries out a longer search (f) it will also approximate the ideal observer in the limit.

More precisely, START learning proceeds as follows (and see Fig. 4d–g): When observing a particular dynamic scene, the learner initializes their interpretation of the scene using the summary statistics, by sampling from their distribution over the various parameters needed to specify a theory. For example, if the summary statistics predict that a particular kind of object has high mass with probability 70% (based on its velocity, say), then there is a 70% chance that the initial hypothesis will assume that object has a high mass. Note that this initialization uses the summary statistics to define a pseudo-likelihood over the theories, rather than mapping directly between summary statistics and physical judgments as in the SSS model. Following this initialization, the learner either remains with their current hypothesis, or proposes random adjustments to it, using the ideal observer model to score the proposed changes and decide whether to accept or reject them. Similar integrations of 'informed' bottom-up initialization followed by a search that is scored by a top-down generative model are a current exciting direction in machine learning (see e.g. Kulkarni, Yildirim, Kohli, Freiwald, & Tenenbaum, 2014; Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). More generally, such a two-system view is similar to dual-process proposals for a rapid, bottom-up system working alongside or in combination with a more effortful and informed system (e.g. Stanovich & West, 2000), although in our case both systems are assumed to be implicit.

In the limit of proposing many adjustments to the theory, this search is guaranteed to eventually converge on the posterior distribution of the ideal observer, as it is implementing MCMC on the search space defined by that observer. In the limit of making no proposed adjustments, this process will on average look just like learning based on the summary statistics. An ensemble of learners terminating their search after different numbers of adjustments – some searching longer, some shorter – will in some sense interpolate between the behavior of the SSS and IO models, but in a way that combines their most appealing properties. The START model essentially initializes its estimate of the theory from the SSS model's inferences, and then iteratively proposes improvements to the theory which are accepted probabilistically to the extent that they improve the Bayesian posterior score – the same criterion for the IO model's inferences, but now computed only on the two candidate theories under consideration (old and proposed) rather than for all theories in the full hypothesis space. START thus integrates the key mechanisms of IO and SSS, combining the efficiency of SSS with the accuracy guarantees of IO.

To complete the description of the START model, we must specify the distribution over learners' search lengths N – how many MCMC samples each learner considers. We consider learners as drawing N from the following distribution:

$$p(N = n) = \alpha \cdot \delta_0^n + (1 - \alpha) \cdot Erlang_2(n, \lambda) = \alpha \cdot \delta_0^n + (1 - \alpha) \cdot \frac{1}{\lambda^2} \cdot n \cdot e^{-\frac{n}{\lambda}}, \quad (1)$$

where δ_0^n is the Kronecker delta, equal to 1 when $n = 0$ and equal to 0 otherwise. This distribution is equivalent to the statement that learners either stay with their initial anchoring (with probability α), or adjust their guess by taking some number of steps N sampled from an $Erlang_2$ distribution with a scale parameter λ (and see Lieder, Griffiths, & Goodman, 2012; Lieder, Griffiths, Huys, & Goodman, 2016 for a similar approach to anchoring using a Poisson distribution). The START model thus has two parameters in addition to the ideal observer or summary-statistics models (which only have noise parameters). Across all physical properties, we set the mixture parameter to $\alpha = 0.5$ (representing an equal probability that the learner will stay with their initial anchoring versus adjust their initial guess through search), and $\lambda = 100$. The resulting distribution is shown in Fig. 4g. The results in the following sections are not highly sensitive to perturbation around these parameter settings. Setting α between 0.25 and 0.75, or λ to 75 or 125, produces qualitatively similar results.

Psychologically, there are at least two distinct interpretations for this bimodal distribution of sampling times, corresponding to models of individual differences in learning style and trial-by-trial behavioral variability, respectively. Both interpretations in some sense reflect the two-step process described above, according to which learners either decide to stay with their initial "gut-feeling" based on the summary statistics (with probability α), or to conduct in addition a slower search through the space of possible explanations for the scene. But in one interpretation, α describes the probability that a learner falls into one of two clusters of learners: those who always stay with their initial guess, and those who always search further to improve it. Under the second interpretation, α describes the probability that any given learner on any given trial will stay with their initial guess, or search further. Our experiments here do not distinguish between these interpretations, but this could be an important question for future work, as we take up in the discussion.

Note also that while the START model might appear to be equivalent to a simple mixture of the summary-statistics and ideal-observer models, this is true only in the limit of very many steps ($N \to \infty$). In our setting, it would require extremely large values of N to accurately estimate the full posterior over the hypothesis space of candidate theories (in general, millions of samples, given that there are between thousands to tens of thousands of hypotheses, depending on the scenario). In practice, however we find that the best fitting versions of the START model typically use values of N around 100. Thus START never comes close to sampling the full hypothesis space, and only succeeds in generating good (i.e., high posterior) samples because of how it uses the feature-based initialization. Rather than viewing START as just a mixture of our other two models, we feel it is more appropriate to view it as a distinct third model potentially combining their two strengths – that is, as a psychologically plausible algorithmic approximation to an otherwise intractable Bayesian learning problem.

3. Experiment

A large-scale web-based experiment was designed to compare our several models of learning physical parameters with people's judgments on a range of "hockey puck" scenarios, the novel dynamic task introduced in previous sections.

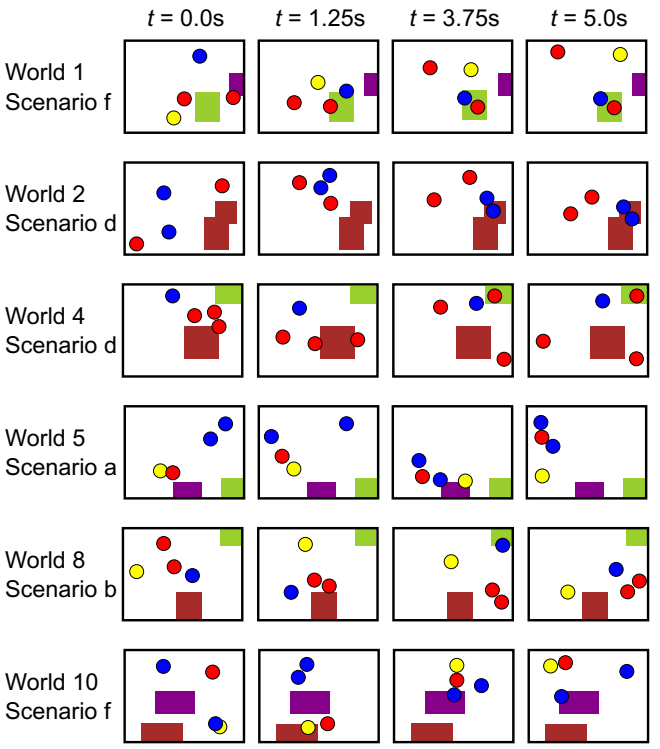

Fig. 5. Snapshots of the dynamic stimuli used for six scenarios out of the total 60. There are 4 images per scenario, showing it unfold over time. The images were sampled at the start of each scenario, 1.25 s into the scenario, 3.75 s into the scenario, and at the end of the scenario (5 s after it started). The scenarios illustrate different physical parameters, some more difficult to see in static snapshots than others. The scenarios illustrate properties such as heavy friction (such as the green surface in scenario 1f, and brown surface in 2d), different mass ratios (for example, blue is heavier than red in 4d), attractive forces (red to blue in scenario 5a, and red to red in 8b), repulsive forces (red and yellow in scenario 8b, red and red in scenario 4d) and global forces (an Eastwards force in 8b, and a Northwards force in 10f).

3.1. 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.

3.2. Stimuli

60 videos were used as stimuli, each lasting 5 s and depicting the dynamics of 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://tomerullman.org/physics-cogpsy-2017/rules.html.

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 third 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.

Using the dynamical rules of the world and starting from the initial conditions, we unfolded the scenarios over 400 steps, and created a video detailing the motion of the objects over time.[4] All stimuli used are available at http://tomerullman.org/physics-cogpsy-2017/stimuli.html, and a static visual representation is shown in the Appendix, in Figs. 11 and 12. Two representative static examples are shown in Fig. 5.

[4] We used the classical Runge-Kutta method (RK4) for numerical integration to move the entities forward in time.

3.3. 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 also that objects of the same color have the same properties. It was explained that objects can be heavy, medium or light, and that each object type can potentially exert forces on other types: object types either attract, repel or don't interact with one another. Participants were instructed to think of the videos as similar to 'hockey pucks moving over a smooth white table-top', and informed that patches on the plane can have different roughness. Finally, they were told there may or may not be a global force in the world, pulling all objects in a particular direction (north, south, east or west). An example experiment with the complete instructions and layout used is available at tomerullman.org/physics-cogpsy-2017/experiment.html.

After the presentation of each video, participants rated the entire set of possible physical properties. Participants were allowed to watch each of the videos as many times as they wanted when answering the questions. For each puck color, participants were asked 'How massive are [color] objects?' Possible answers were 'Light,', 'Medium,' 'Heavy,' or 'Can't tell from movie.' For each surface color, participants were asked 'How rough are [color] patches?' Possible answers were 'As smooth as the table-top,' 'A little rough,' 'Very rough,' or 'Can't tell from movie.' For each puck color-pair combination, participants were asked 'How do [color 1] and [color 2] objects interact?' Possible answers were 'Attract,' 'Repel,' 'None,' or 'Can't tell from movie.' Finally, participants were asked 'Is a global force pulling the objects, and if so in what direction is it pulling?' Possible answers were 'Yes, it pulls North,' 'Yes, it pulls South,' 'Yes, it pulls East,' 'Yes, it pulls West,' or 'No global force.' This gave us a total of 13 questions per video, and 5 videos gave us a total of 65 data points per participant. The 'Can't tell from video' answer was supplied for cases where the question is not relevant, for example a question regarding the mass of blue pucks when no blue pucks are shown in the video.

3.4. Results

We analyzed the results in two ways, one looking at overall performance levels on all questions, and one looking at more fine-grained response patterns on specific questions.

3.4.1. Overall performance

Participants correctly answered 54% of the questions on average, with a standard error of 13%. This correct response rate is highly above the chance performance of 32% as shown in Fig. 6.[5]

Participants' performance differed depending on the particular physical property being considered. The correct response rate broken down by property was: 43% for mass, 44% for friction, 62% for pairwise force, 68% for global force. The chance performance is 33% for all properties except global force, where it is 20%.

The correct response rate for mass and friction appears low, but this basic analysis does not take into account cases where participants saw only two types of pucks or patches and correctly ordered them. For example, suppose a participant saw only yellow and red pucks in a video, and responded 'red: light, yellow: medium', while the ground truth was actually 'red: medium, yellow: heavy.' The basic analysis would count this as two wrong answers for the participant (wrong answer on yellow, and wrong answer on red). If we consider correct orderings to be correct answers as well, participants' performance increases to 48% correct on mass and 66% correct on friction (compared with 33% chance performance in both cases). The overall correct response rate when taking such cases into account is 59%.

3.4.2. Fine-grained patterns of performance

Aggregating over the different scenarios: We obtained the empirical distribution of responses over the possible answers across all scenarios, and collapsed across the property of color to consider four 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, thus we can fit an ordinal logistic regression to the participant data, shown in Fig. 7a. The figure displays the cumulative probability on the y-axis, and the relevant response is color-coded according to the label. For example, on this regression the probability people will answer 'heavy' when the true mass is in fact light (equal to 1) is 15%. The probability that they will answer medium is 33% (48% minus 15%), and the probability that they will answer 'light' is the remaining 52%. This is close to the empirical values of 16%/37%/47%.

An ordinal regression cannot be used for the global and coupling forces, and so Fig. 7c shows empirical confusion matrices, detailing the percentage of people that chose each option given the ground truth.

[5] The exact number of potentially correct questions varied by world and scenario, as some questions were not relevant for some stimuli, such as a question about the mass of blue pucks when no blue pucks were shown. The chance performance distribution was calculated by simulating 20 separate experiments of 290 participants. Each simulated participant in each simulated experiment was paired with a real participant, but chose at random for any given answer by the real participant.

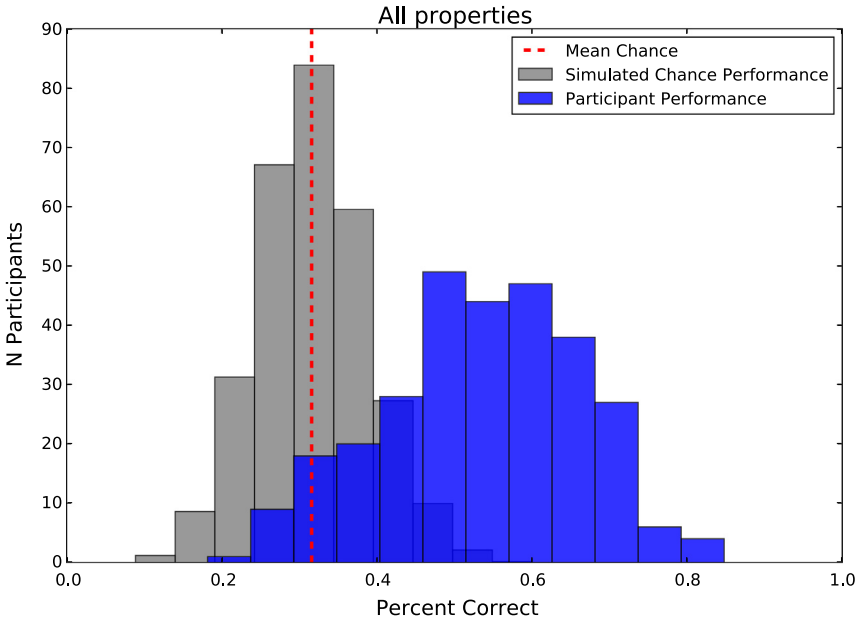

Fig. 6. Distribution of correct response rate across all physical properties compared to chance. The blue bars show the participant distribution, while the gray bars show the expected distribution if participants were guessing at chance. The dotted red line shows the mean of chance performance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

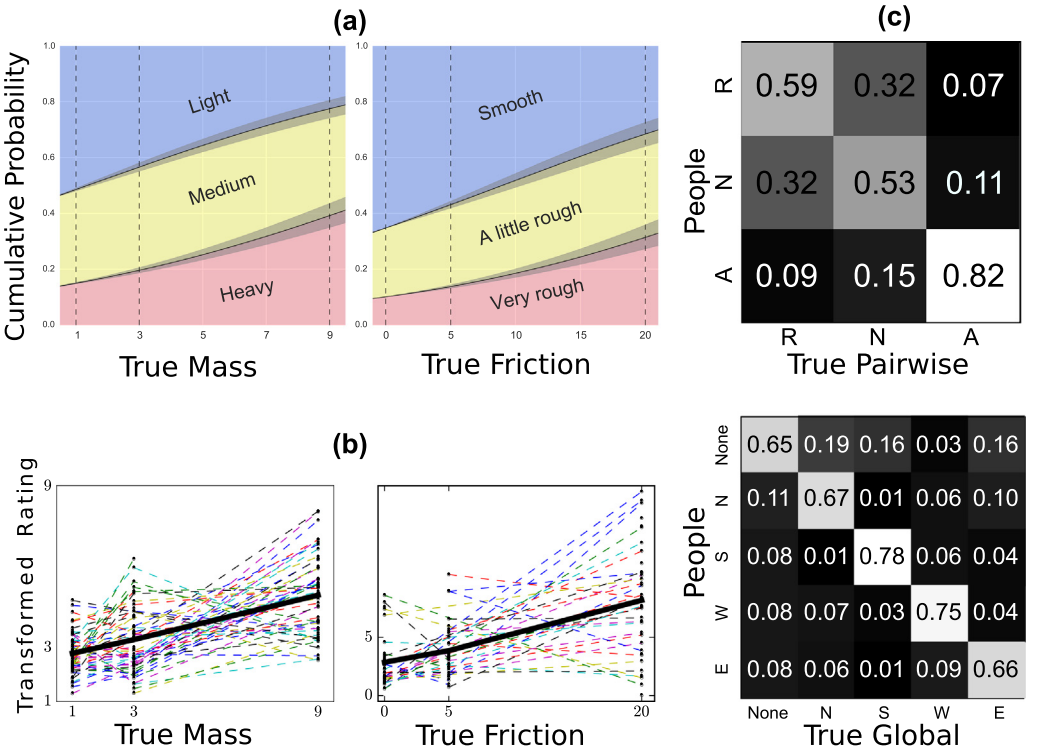

Fig. 7. Analysis of participant performance using: (a) Ordinal logistic regression for mass (left) and friction (right). Shaded black areas represent uncertainty on parameter estimates, colored patches show the ordinal responses. The upward trend indicates a greater proportion of participants selecting 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). The labels R, N, and A in (top) stand for Repulsion, No Force, and Attraction. The labels N, S, W, and E in (bottom) stand for North, South, West, and East. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Transforming responses per scenario For mass and friction we can assess participant performance in a more refined way, by considering the distribution of responses for each puck (and surface) in each one of the 60 scenarios, and transforming this distribution into a quantitative prediction for that puck (or surface). We do this by taking 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 participant rating is $0.6 * 9 + 0.4 * 3 = 6.6$), and comparing the results with the ground truth, shown in Fig. 7b. These sub-figures plot the average rating of participants for mass/friction in a given scenario, compared to the 'ground truth.' Each black dot thus represents the average rating of 25–30 participants for mass/friction. The black solid line shows the average response for all masses across all scenarios. Dotted colored lines connect masses/friction in the same scenario, thus a rising line means a correct ranking. We next consider each property separately.

3.4.3. Results by physical property

Mass: The upward trend of the lines in the logistic regression, shown in Fig. 7a, shows that participants correctly shift in the probability of answering that a mass is heavier when that is in fact the case. The linear correlation depicted in Fig. 7b shows that although there is a large degree of variance for any given mass, participants were able to overall correctly scale the masses.

Friction: Again we see a upward trend in the logistic regression, shown in Fig. 7a, again indicating correct sensitivity to the underlying friction. 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.

Pairwise forces: As shown in Fig. 7c 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.

Global forces: As shown in Fig. 7c 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 a null-force than it is with any other force.

3.4.4. Between-participant agreement

The previous analysis gives us a measure of participant performance and variance with respect to the ground truth. While participants might stray from the ground truth, they might also be in agreement about how to stray. We can assess within-participant agreement by measuring the split-half correlation of participants' parameter estimations for the different scenarios (for the ordinal cases of mass and friction), or the split-half correlation between the probability judgments (for the nominal cases of pairwise and global forces). We ran through this process several thousand times to obtain confidence intervals on these correlations, with results summarized in Table 1.

These values show that participants are largely in agreement with one another. The values also place a useful upper-bound on what can be achieved with a given predictive model.

3.4.5. Discussion

Participants' overall performance when compared to the ground truth is far from perfect, but it is impressive given that they had to estimate multiple parameters from just a few seconds of passive viewing. Furthermore, there is general between-participant agreement, indicating that if participants are deviating from the ground truth they are doing so in a systematic manner. There are several points of interest for patterns of performance on each physical property:

For mass, the apparent ability to correctly rank and quantitatively scale multiple masses is of particular interest, as experiments on inferring mass from collisions have usually focused on judgments of mass ratios for two masses, often requiring binary responses of more/less massive (e.g. Gilden & Proffitt, 1989).

For friction, the linear correlation depicted in Fig. 7b 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.

For pairwise forces, in addition to clearly better performance on attractive forces, there are interesting asymmetries in how people think about the absence of forces. Participants tend to confuse "no pairwise forces" with repulsion much more

Table 1
Between-participant correlations for each physical property, with bootstrapped 95% confidence intervals.

	Mass	Friction	Pairwise	Global
r	0.72 ± 0.07	0.82 ± 0.04	0.87 ± 0.04	0.89 ± 0.03

than they confuse it with attraction, and also more frequently choose "no pairwise forces" when repulsion is present than when attraction is present. We will return to these points in the next section, as they point to some of the features that people may be using to judge pairwise forces that treat attraction very different from repulsion.

For global forces, if participants did not correctly interpret the display as shown from a 'bird's eye view', then the 'South' direction could be interpreted as 'Down' and so activate certain prior expectations about a gravity force pulling in that direction. While this was indeed the most correctly perceived force, it is a small effect, and such an explanation does not account for why a force pushing West, for example, is better detected than one pushing East.

3.5. Analysis of stimulus features

In addition to asking about people's accuracy and consistency in estimating physical parameters in our tasks, it is of interest to identify which stimulus features they are most sensitive to or most likely using to guide their judgments. This is important both for grounding the choice of stimulus features used in our summary-statistics based models (SSS and START), as well as more generally for understanding the relative importance of different aspects of object motion for how people learn various physical properties and force relations from dynamic scenes.

We conducted a regression analysis to answer these questions. We used all of the stimulus features listed in Section 2.2, and regressed participants' choices on the relevant features for each kind of physical parameter, in order to assess the relative weights of each feature for each kind of judgment. In order to make a direct comparison between the regression weights, we z-transformed the feature-data of the different scenarios. We then used a multinomial logistic regression on participant response with the features as contributing factors (ordinal regression for mass and friction, and nominal regression for pairwise and global forces). Table 2 summarizes the resulting regression weights, and indicates the probability that they are significantly contributing to the regression (under a χ^2 test).

The results of this 'model-free' regression indicate that people are likely sensitive to most of the features considered. Beyond statistical significance, we find that some features are more important than others, as measured by their relative weight. For mass, while much previous research (e.g. Gilden & Proffitt, 1989; Runeson et al., 2000) has focused on the post-collision angle and velocity ratio, here we find that simply the average velocity of an object is a better indicator of whether people will judge it as heavy or light (faster pucks are seen as lighter), although this cue is not highly weighted relative to the cues for the other features. For friction, the time an object spent at rest on a surface was more important than its deceleration profile, which is surprising given the saliency of a puck slowing down rapidly. For attractive forces, average distance carries more weight than the change in that distance. For global forces, the average position along the x-axis was more highly weighted on average compared to the change in the average x-axis position, but the reverse was true for the y-axis.

Thus, the general pattern of regression weights within and across properties suggests that people might make more use of positional features relative to movement features, and that they prefer simpler motion signals (lower-order derivatives, such as velocities) to more complex ones (higher-order derivatives, such as accelerations, or combinations such as velocity or angle ratios) when both are available. The y-axis features for the global forces deviated from this overall pattern, but such a deviation may be due to the regression over-fitting to the relatively small number of scenarios that contained active global forces.

The uniformly low regression weights for repulsive pairwise forces, in contrast to the very strong weights for attraction, might also suggest a basis for the asymmetry we found in how people estimate pairwise forces. Recall that people were

Table 2
Logistic regression weights, by physical property and relevant feature. Asterisks indicate statistical significance at the $p < 0.05$ (*), $p < 0.01$ (**) and $p < 0.001$ (***) levels.

Mass	Average Velocity	Velocity Ratio	Angle Ratio
Weight	0.37***	0.11***	0.18***

Friction	Deceleration	Zero Velocity
Weight	-0.31***	−0.64***

Pairwise Force	Average Pairwise Distance	Δ Pairwise Distance
$Weight_{Repel}$	0.14***	0.13***
$Weight_{Attract}$	−1.78***	−0.63***

Global Force	Average X	ΔX	Average Y	ΔY
$Weight_{North}$	−0.14	0.26*	−0.45***	−1.14***
$Weight_{South}$	−0.25	−0.18	1.09***	0.48***
$Weight_{West}$	−1.02***	−0.5***	0.38*	−0.55***
$Weight_{East}$	0.72***	0.65***	0.06	−0.18

much more likely to confuse no pairwise force with repulsion, as opposed to attraction. The regression analysis suggests there is just not much signal in the stimulus features we can easily identify that uniquely picks out repulsion, while in contrast there are very strong signals that pick out attraction, most notably the simple property of average relative position which is the most highly weighted of all features in our analysis. To put this in plain terms, when two objects attract each other in our scenarios, they tend to spend much of the movie (especially the later few seconds) very close to each other. This makes average pairwise distance a reliable cue for attraction. In contrast, if two objects repel each other, then in our scenarios they tend to stay far from each other over the course of the movie, and there is little change in their relative position. Because of how much else is typically happening in these complex scenes, this is also the case if they have no pairwise forces between them. Hence the aggregate position and movement statistics we are considering, which in most cases are useful features for observers to consider, help relatively little in distinguishing repulsion from no pairwise force, though they are very distinctive for attractive forces.

The regression weights can be used to predict participants' parameter estimations for the different scenarios (for the ordinal cases of mass and friction), or the probability judgments (for the nominal cases of pairwise and global forces). Doing so results in a correlation of $r_{mass} = 0.77, r_{friction} = 0.78, r_{pairwise} = 0.82$ and $r_{global} = 0.91$. It is important to keep in mind that this regression model has a large number of free parameters, including the weights and intercepts (5 free parameters for mass, 4 for friction, 6 for pairwise forces and 16 for global forces) and thus may be over-fitting. We performed cross-validation to account for this, regressing on half of the participant data (training set) and testing the correlation when predicting the half left out (test set). This leads to an unsurprising drop in correlation, summarized in Table 3.

The above regression analysis can be seen as a purely feature-based, theory-neutral approach to modeling how people estimate physical parameters. It is not on its own a theory of how people could learn about physics: the regression model has 31 free parameters (including weights and offsets), and we have no reason to expect that the best parameter settings found here will generalize to settings beyond those we have studied. It does however establish a useful benchmark for assessing how well the theory-based learning models we focus on – models which can directly generalize to new worlds and scenarios – capture human judgments in the tasks we study here. We return to these models in the next section.

The regression analysis also provides qualitative insight into the cues participants attend to in our scenarios, and might use in similar scenarios. While we can make some general conjectures about the relative weights of different kinds of features, the regression on its own does not provide a rationale for why the features should have the weights that they do. The relative magnitude of the weights can be analyzed in a more systematic manner, by considering how *diagnostic* each feature is for a given setting of a physical property, and assessing whether people's sensitivity to that feature (measured as the magnitude of the regression weight) corresponds to how diagnostic the feature is. For example, observing that the particles are on average drifting to the left might be highly diagnostic of a West-directed global force, but not very diagnostic of a North-directed force.

As a measure of how diagnostic a feature is, we use the expected amount of information gained by observing the feature's different values (for a similar treatment see Oaksford & Chater, 1994). That is, we begin by assuming a prior distribution $p(V)$ over each physical property's values V_i, with an uncertainty $I(V_i)$ corresponding to:

$$-\sum_i p(V_i)\log(p(V_i)). \tag{2}$$

For simplicity, we assume an initially uniform prior distribution on the property values. Once a feature F is observed to have a particular value f, it changes the probability distribution over the values of the physical property to a posterior distribution $p(V|F=f)$, calculated using Bayes' theorem:

$$p(V_i|F=f) = \frac{p(F=f|V_i)P(V_i)}{\sum_i p(F=f|V_i)P(V_i)}. \tag{3}$$

This posterior distribution results in a new uncertainty $I(V_i|F=f)$:

$$-\sum_i p(V_i|F=f)log(p(V_i|F=f)). \tag{4}$$

We can then define the information gained from observing the feature as:

$$I_g = I(V_i) - I(V_i|F=f). \tag{5}$$

However, it is not known in advance what particular feature value will be observed. We can thus consider the *expected* information gain:

Table 3
Cross-validated correlations between participant predictions, and the predictions of a multinomial logistic model that was regressed to participant answers.

	Mass	Friction	Pairwise	Global
r	0.72	0.67	0.74	0.85

$$E(I_g) = I(V_i) - \sum_f p(f) I(V_i | F = f). \quad (6)$$

While $p(F)$ does not have a closed-formed solution, we can estimate it for each physical property by conditioning on a particular property value, and sampling many dynamic scenarios given that value. We can then calculate the features values for these different scenarios, similar to the process that was carried out for the SSS model (but for a different purpose). Essentially, the procedure is:

1. Set a given physical property value $V = V_i$.
2. Sample many dynamic scenarios conditioned on that property value.
3. Calculate feature F for the resulting scenarios.
4. Calculate the information I_g for each $F = f$, assuming an initially uniform prior.
5. Take the mean I_g over the sampled scenarios.

This procedure gives us an estimate for how diagnostic a feature F is for a property value V_i, for example how diagnostic Δx is for detecting a Westward force.

We carried out this process for each property and feature detailed in Table 2, and found a correlation of $r = 0.57$ ($p < 0.01$). Fig. 8 shows the correspondence between the magnitude of the regression weights and the expected information gain. This correspondence shows that people's sensitivity to a feature (estimated by the regression weight) is in quantitative agreement with how diagnostic that feature is for a given physical property value. To summarize, this analysis shows that people are indeed sensitive to the features considered, independent of any particular model we later use. While a feature-based regression model on its own cannot provide a rationale for the different feature weights, we found that the degree of participant sensitivity can be partly explained with reference to how diagnostic a feature is, lending support to the view that participants have access to the possible distributions of these features.

4. Comparison to models

We now turn to a comparison of participants' performance with the theory-based models introduced previously, the IO, SSS, and START models.

For the *Ideal Observer* model (IO), we obtain predictions in the following way: For each scenario, we fix the observed initial conditions and simulate the resulting paths for all the relevant physical theories. We then give each theory a log-likelihood score by assessing the deviation of its simulated path from the observed path. Finally, for each parameter of interest we marginalize over the other parameters by summing them out, to obtain a log-likelihood score for the relevant parameter.

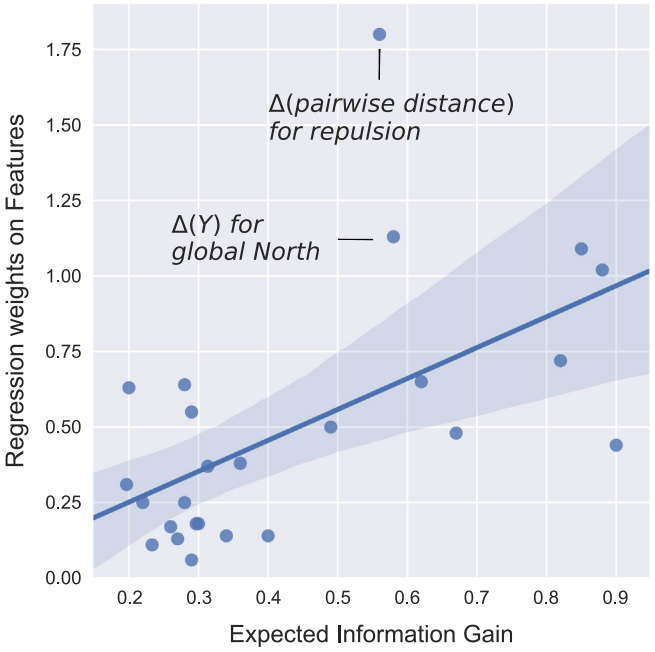

Fig. 8. Correlation between expected information gain for the different features per physical property, and the magnitude of the regression weight of those features. Shaded areas show bootstrapped 95% confidence intervals. Two points in particular are highlighted, as examples of features that are weighted more than predicted by the expected information gain measure.

For the *Simulation and Summary Statistics* model (SSS), we obtain predictions by following the procedure detailed at the end of Section 2.2, and illustrated for a particular example in Fig. 3.

For the *START* model, we obtain predictions by following the procedure detailed at the end of Section 2.3, modeling the participants as an ensemble of learners that guess an initial setting of the physical parameters by sampling from the SSS model, and then searching for a varying amount of search-steps in the space of possible parameter settings.

These different parameter estimates result in predicted distributions over the responses for each physical property, for each scenario. We begin by collapsing across scenarios so that we can compare the results to the logistic regressions and confusion matrices of the participant data shown in Fig. 7a and c. Note that each model includes a free 'noise' parameter applying to the distributions across all scenarios, which allows us to bring each model as close as possible to the participant data. We consider 'close' as minimizing the RMSE[6] between the different distributions of the empirical confusion matrices (for pairwise and global forces) or the confusion matrices predicted by the logistic regression (for mass and friction).

We begin by considering the ordinal logistic regression as applied to the different models, compared with mass and friction, shown in Fig. 7. For mass inference, the SSS model slightly outperforms the IO model (when summing across the mass range). The START model is at a mid-point between the two, indistinguishable given the confidence intervals. For friction inference, the IO model outperforms the SSS model in terms of how close it is to people's judgments (when summing across the friction range), although here START outperforms both.

We next consider the confusion matrices. Of particular interest is the confusion matrix for pairwise forces, where people showed an asymmetry in their confusion of the absence of force. That is, when there actually is an absence of a pairwise force, people incorrectly rate this as a repulsive force much more than they incorrectly rate this as an attractive force (32% repulsive compared with 15% for attractive, see Fig. 7c). As mentioned in Section 3.4.5 on feature analysis, we can understand this difference intuitively: an attractive force is more likely to pull bodies closer together, which makes the attraction stronger and so gives further evidence for the attractive force. A repulsive force pushes bodies further apart, growing weaker and providing less evidence for its existence over time. But such an asymmetry plays out over the entire dynamic scene. This asymmetry does not come naturally out of the IO model, which sums up the error along local deviations between a simulated trajectory given by a particular theory, and the observed trajectory. In such a model the local error produced by a theory that posits an attractive pairwise force is the same as that produced by a theory that posits a repulsive force.

By contrast, a summary statistic looking at the average pairwise distance does replicate this asymmetry. As illustrated in Fig. 3c, when we condition on the absence of force (in blue) and on a repulsive force (red), we generally find an overlap in the distribution of the summary statistic that is greater than that between the absence of force and an attractive force (green). The START model reproduces a confusion matrix that is similar to people's performance, shown in Fig. 7c. In particular, we reproduce the asymmetry between repulsion and the absence of a pairwise force (27% repulsive compared with 16% for attractive). While this asymmetry also exists for the SSS confusion matrix, the START confusion matrix is closest to that of people.

The second confusion matrix to consider is that of global forces. As mentioned, for people one of the main points of interest was the confusion between any given force and the absence of force, relative to any other force. While the IO seems to replicate this finding, the SSS model does not. Also, we interestingly find that the SSS model is quite bad at detecting the absence of global forces, perhaps because none of the simple features we used account for a null-force. We take up the question of other possible features, including more force-based ones, in the discussion. The START model produces a confusion matrix that is as close to people as the IO model, including the apparent confusion between any given force and the absence of force.

To sum up these results: comparing accuracy and error rates of models and people, we found that all three models fit mass estimates equally well; the START model fits friction judgments slightly better, and pairwise force estimates substantially better than the other models; and both START and IO models fit global force judgments equally well, and better than the SSS model. Only the START model captures all the interesting qualitative asymmetries in people's estimates of pairwise forces: that attractive forces are more accurately identified than either repulsive forces or "no pairwise force", and that "no pairwise force" is more frequently mistaken for repulsion than attraction, and a more frequent incorrect choice when repulsion is present than when attraction is the true state.

Having examined the aggregate results, we can refine our comparison by looking at the response distributions the models give in each scenario and for each object and property, correlated with those of people. For mass and friction coefficient judgments, we can compare between participants and the different approaches by again converting posteriors into predicted mass and friction values. For global and pairwise forces we can compare performance by correlating the predicted model posteriors for each scenario and property, with the posterior as calculated from normalized participant judgments.

The comparison of these various approaches with people is summarized in Table 4 below, showing correlations between people and different approaches. Note that the 'noise' parameter mentioned earlier could be used to optimize this linear correlation, but in order to reduce the number of free parameters we re-used the noise obtained from the previous comparison. We used a standard bootstrap method to obtain estimated confidence intervals on these correlations (Efron & Tibshirani, 1986).

[6] We also considered using KL-divergence as the distance metric, but that does not alter the results.

Table 4
The correlation between people's judgments of different physical properties and the different computational approaches: Ideal Observer (IO), Summary Statistics Approximation (SSS), and START. Correlations include 95% estimated confidence intervals, calculated using 10,000 bootstrap samples.

	Mass	Friction	Pairwise	Global
IO	0.57 ± 0.08	0.54 ± 0.13	0.55 ± 0.04	0.89 ± 0.02
SSS	0.59 ± 0.13	0.68 ± 0.11	0.69 ± 0.04	0.82 ± 0.03
START	0.59 ± 0.08	0.67 ± 0.10	0.75 ± 0.03	0.90 ± 0.02

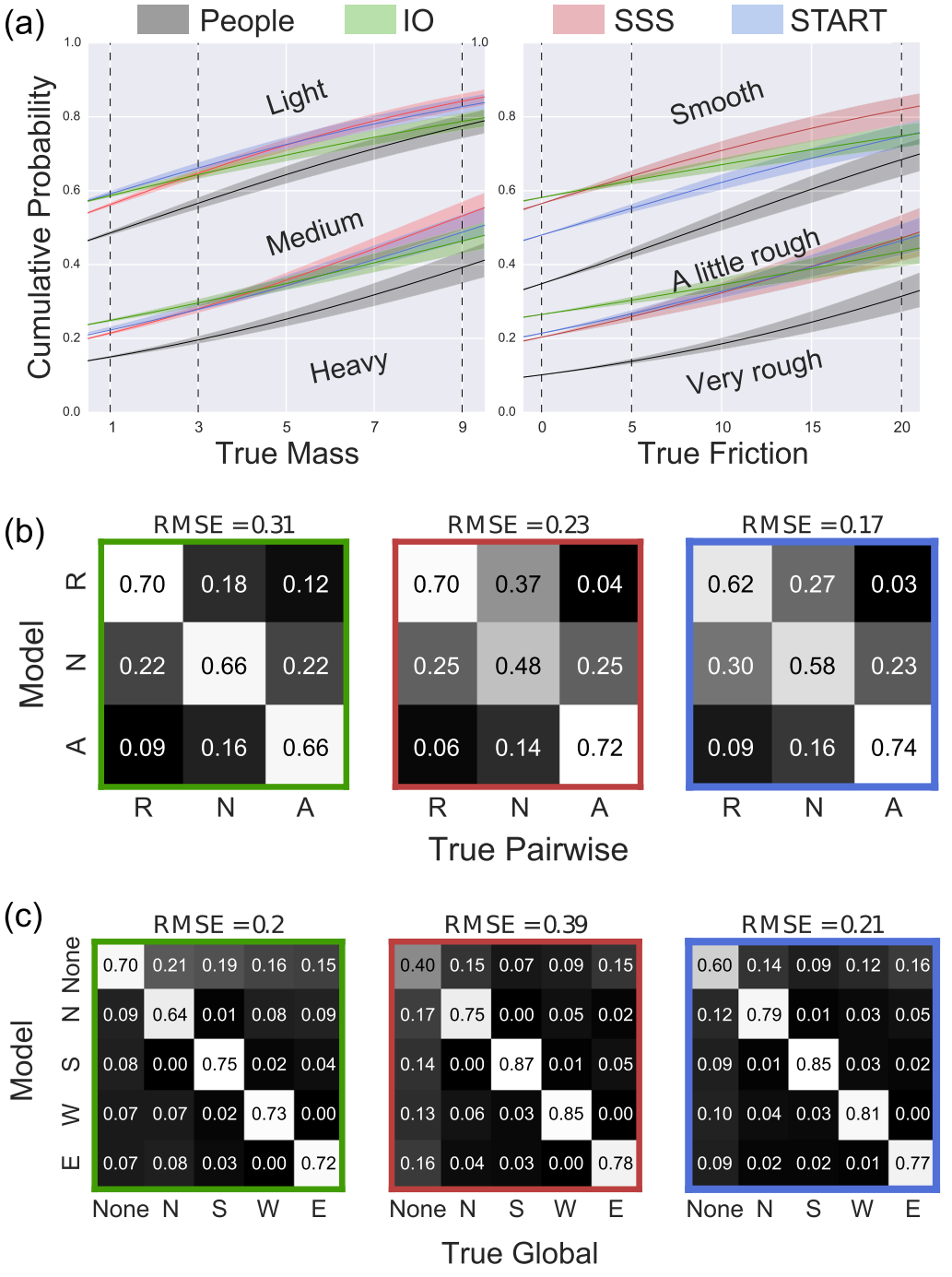

Fig. 9. Comparison of different model performance for properties (a) friction and mass (b) pairwise forces and (c) global forces. The labels R, N, and A in (b) stand for Repulsion, No Force, and Attraction. The labels N, S, W, and E in (c) stand for North, South, West, and East.

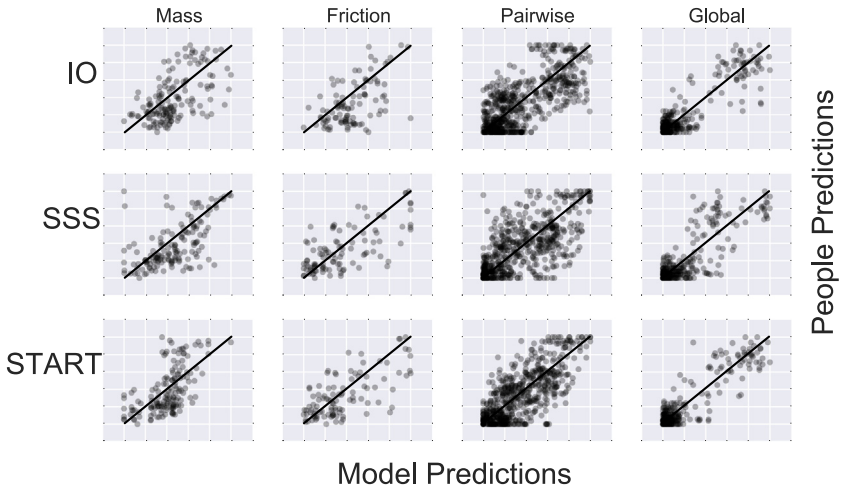

Fig. 10. Correlations between people's answers and those given by the different models, for the four physical categories. Model predictions are on the x-axis while participant prediction are on the y-axis. Each data point represents the average of participants' responses for a single parameter value in a single scenario (e.g., the mass of the blue objects in World 5, Scenario a), as well as the mean of the corresponding model's predictions for that same parameter value.

As can be seen from Table 4, the models correlate similarly with human judgments across all four types of parameters, but with several significant differences. The START model consistently performs as well as or better than either the IO and SSS models across all four judgment types: All models correlate equally well with humans' mass judgments; the SSS and START models perform slightly but not significantly better than the IO model on friction judgments; the START model correlates better than either alternative model on pairwise force judgments, while both START and IO models beat the SSS model in predicting global force judgments. This picture is similar to the overall results described above for accuracy and errors, shown in Fig. 9. In Fig. 10 we illustrate these correlations in more detail with scatter plots.

The START model is also similar in performance to the pure feature-based regression (Table 3). This feature-based regression uses about 5 times the number of free parameters that are fit heavily to participants' judgments, but obtains correlations that for all four judgment types are not significantly better than the START model's. This suggests that START's smart initialization through summary statistics followed by short periods of stochastic search guided by the Bayesian theory score, may be close to capturing the predictable variance in what humans can learn about physical parameters from brief dynamic scenes such as the ones explored here. However, the comparison to the benchmark of between-participant agreement (Table 1) suggests that the model still has room to improve.

5. General discussion

Taken together, our experimental and modeling studies suggest that a combination of hierarchical Bayesian learning over a space of probabilistic programs and efficient feature-based approximate inference algorithms offers a useful framework for explaining how people can learn aspects of intuitive physics from observations. Our experiments presented a challenging learning task: Observers judged up to thirteen different dynamic aspects of scenes – objects' relative masses, friction coefficients, and both global and pairwise forces for three types of objects – after observing just a few seconds of the objects in motion, with perhaps a few stimulus repetitions. Even in our simple "hockey puck" domain, the joint hypothesis space of physical theories that could explain a particular observed scenario contains upwards of ten thousand hypotheses. The central computational challenge raised by these studies is to explain how people can learn such rich physical theories from so little data, so quickly. Each of the several components of our theoretical framework plays a crucial role: Probabilistic programs provide a rich language for representing these hypotheses and the patterns of motion they predict, hierarchical Bayes provides a framework for learning these theories from sparse data, and smart initialization of stochastic search in the space of theories guided by summary statistics of objects' motions provides an approximate inference scheme that rapidly converges on high-posterior probability theories.

We found that people performed relatively well on this learning challenge in objective terms, and were generally consistent with one another. These findings show that people's ability to infer physical parameters underlying object dynamics is not limited to inferring single physical parameters from very simple isolated events (such as inferring relative masses from a single collision between two objects), as studied in previous laboratory work, but instead generalizes to more complex scenes more representative of the real-world challenges of learning physics.

The fact that participants' performance was well-described by computational models that performed probabilistic inference over physics-engine-like simulations stands in contrast to previous computational models for inferring physical parameters defined in terms of feature-based heuristics (Gilden & Proffitt, 1989; Todd & Warren, 1982) or fixed-structure Bayesian

networks (Sanborn et al., 2013). Previous models either do not generalize to our settings, or explain only a small fraction of the explainable variance. Our studies thus provide further evidence for probabilistic inference in physics-engine-like simulations as a general framework for modeling how people perceive and predict the physics of complex scenes (e.g. Battaglia et al., 2013; Hamrick et al., 2016; Smith & Vul, 2013), and show how these accounts can be extended to model how people learn the parameters of physical theories from observing dynamic interactions.

People also made systematic errors in our tasks, suggestive of how they might be using bottom-up summary statistics to approximate Bayesian learning in a computationally efficient and cognitively plausible manner. In particular, we found that the START model emerged as the best account of participants' learning behavior, based on both overall quantitative fit. as well as the model's ability to fit specific patterns of systematic error. The START model's general approach of initializing a short MCMC search smartly from the diagnostic summary statistics also makes good engineering sense: It can transcend limitations of either the ideal Bayesian observer or the summary statistics approximation on their own, 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. Even with algorithmic approximations to the ideal observer, the search might take a large number of samples if initialized randomly. The feature-based statistics are useful heuristics in many cases, but are unable to handle situations that deviate from the norm.[7] Also, summary statistics in our setup do not replace the knowledge of a generative model, since they require simulations from a generative model in order to estimate their distribution. The computational intensity of the full ideal model is not as much of an issue for the START model, as it only considers a limited number of steps in a theory-space search.

The START model makes the assumption that people either stay with their initial guess with a certain probability α, or adjust their guess by considering alternative samples from a theory-space of explanations. It would be interesting to test whether these two modes of sampling could predict different modes in participant response times and corresponding speed-accuracy tradeoffs. Would there be a subset of trials in which participants respond more quickly and make inferences more correlated with the SSS model, as if they are only using their initial guess, and a complementary subset in which their responses are slower, more variable, more accurate, and more correlated with the predictions of the START model when it performs iterative search to adjust its initial guess?

Further, as mentioned when introducing the START model, the α parameter could represent either a division of participants into two groups (those who maintain their initial response, and those who adjust it), or the possibility that for any given physical inference, each individual has a probability of α of staying with their initial guess rather than performing a longer search. Response times might also be able to distinguish between these possibilities: The separate-populations hypothesis would suggest a bi-modal distribution of response times across participants, in a way that also corresponds with whether participants' inferences are better fit by the START model with few samples or with more samples.

The START model was in part motivated by the fact that exact Bayesian inference over the space of physical parameters quickly becomes intractable as the space of physical parameters and their possible settings grows larger. The START model provides a psychologically plausible sampling-based approximation to the Bayesian posterior, motivated in part by Approximate Bayesian Computation methods for making inferences in complex simulation-based models (Blum et al., 2013), as well as by the general class of sampling-based approaches to implementing Bayesian models of cognition in the mind and brain (Gershman et al., 2015; Griffiths et al., 2012). However, sampling-based schemes are not the only option for approximating challenging Bayesian inferences. One common alternative in machine learning is variational inference (Jordan, Ghahramani, Jaakkola, & Saul, 1999), and our results do not provide specific evidence for a sampling-based approach over a potential variational-based method. This would be a worthwhile question for future work to explore.

Up until recently there were engineering reasons to think sampling-based approximations are more plausible for learning intuitive physics, over and above the recent history of successful sampling-based models in other areas of perception and cognition. Variational approximations to a Bayesian posterior rely on factorizing the joint posterior distribution into a product of independent terms for different partitions of variables, each of which is assumed to come from a family of approximating densities that are more amenable to exact analysis. In practice, deriving useful variational approximations for a given model is not a trivial issue, and it used to require significant domain-specific expertise and engineering. However, automatic variational inference for general probabilistic programs is a topic of current research interest in the AI community (Kucukelbir, Ranganath, Gelman, & Blei, 2015; Ritchie, Horsfall, & Goodman, 2016; Wingate & Weber, 2013), and it would be interesting to see if these lead to computationally and cognitively plausible alternatives to sampling-based methods for learning in probabilistic physics engines. A START-like model might also apply in such a setting, in the sense of using a good initialization and then carrying out a limited adjustment process. One option would be to use the summary statistics to set the initial parameters of the variational distribution, and then to carry out a limited number of optimization steps.

The fact that the Ideal Observer model performed better than the Summary Statistics Simulation model on some properties might be due to other unaccounted-for features that, when used correctly, would bring the SSS model closer to people's performance for those properties as well. In particular, given the relation between forces and acceleration, it might be that including more acceleration-based features would improve performance on force-related inferences, although it is not so clear, as we found that people tended to be more sensitive to lower-order motion features (average positions, rather than

[7] For example, consider a scenario involving two attracting pucks that begin in full contact, rotating around one another and moving together when one is struck. A normally useful statistic for detecting attraction – the difference between the initial and final distance of the pucks – would be useless here. The ideal observer and presumably people would have no problem detecting attraction in such a case.

changes in position), even for forces.[8] We also computed the distributions over the summary statistics by sampling many simulations from the generative model, raising the question of when and how these simulations are carried out. As mentioned in Section 2.2 (and illustrated in Fig. 3), taking only several hundred such simulations provides a reasonable approximation to the limit-case distribution. We speculate that different time-scales might be involved in this process. A small number of simulations might be carried out on the fly while performing inference, but other simulations might continue to take place off-line, and further inform relevant features. Being repeatedly exposed to a particular dynamic environment (whether a new video game or a different physical environment in the real world) can thus cause an increase in off-line simulation, accounting perhaps for findings such as the *Tetris effect*, in which people who play the computer game Tetris later report "seeing" Tetris imagery when not playing the game, including in their sleep (Stickgold, Malia, Maguire, Roddenberry, & O'Connor, 2000).

We used a set of plausible summary statistics based on simple, easy-to-compute features of motion trajectories, many of which have been proposed in the past as the cues driving basic physical inferences (e.g. Gilden & Proffitt, 1989, 1994). However, this set of features is not meant to be exhaustive, not for the set of scenarios we describe here nor for scenarios that differ from them. Reasoning over scenarios that differ in their underlying physics from the ones described here will likely require additional features, but we expect our original list to stay broadly relevant for such new scenarios, as many of the features are basic and generally applicable.

Beyond the particulars of what features to use and the ways to combine them with a generative physics engine, there has been recent debate surrounding the extent to which any kind of mental simulation underlies people's ability to grasp the physics of dynamic scenes (Goodman et al., 2015; Marcus & Davis, 2013). People appear to perform some kinds of physical inferences in ways that suggest a rich mental simulation ability, but not every kind of reasoning that an intuitive physics engine could support is easy and natural for human beings.

Part of the difficulty in some tasks may reside not in the process of simulating physics but in the process of hypothesis generation. Recent work asking people to posit configurations of hidden objects and forces to explain otherwise-anomalous dynamic scenes suggests that naive observers have a hard time generating the best hypotheses to explain trajectories, although they can correctly evaluate them, in ways consistent with running physical simulations forward (Carroll & Kemp, 2015). A similar effect could be at work in our studies here: With an effective hypothesis space of over a hundred thousand possible ways physics could operate in our task, the hardest challenge in learning is probably not evaluating hypotheses – simulating their physical predictions forward and comparing with the observed scene – but rather in effectively searching through the hypothesis space to find plausible candidate hypotheses to evaluate. While the START model we propose represents a first step at addressing this search problem, with a smart feature-based initialization followed by a short run of MCMC to improve this guess, much more work is needed in this direction. We see the current START model as only a good "initialization" that can surely be improved upon. In particular, both the initialization step and the search process for improving hypotheses can likely be made both more psychologically faithful, and more algorithmically efficient. We find it noteworthy that the simple strategies for approximating Bayesian learning embodied in the START model dovetail with recent exciting proposals in machine learning and computational vision (e.g. Wu et al., 2015), as well as with rational process models for explaining biases in classic cognitive judgment tasks (Lieder et al., 2012). We expect that similar models for approximating Bayesian inference in computationally tractable and cognitively plausible ways will come to be increasingly important in both artificial intelligence and cognitive psychology over the next few years.

There are many questions that are still open when considering the challenge of inferring physical dynamics from perceptual scenes. In the rest of the discussion we consider several of these questions, and how our framework might shed light on them.

First, to what extent are the computational processes underlying intuitive physics shared between adults and children? How and when do people learn at the higher levels of the framework? Certain basic physical expectations develop over the first few years of life (Baillargeon, 2002; Needham & Baillargeon, 1993), expectations that may be recast as learning over-hypotheses over forces and properties at higher levels of the proposed framework, for example generating an elasticity property, or tuning a prior over possible mass distributions. Such 'proposals' could be carried out as an algorithmic search over theory space, similar to the one discussed here, though at a more abstract level of physical theories. On the empirical side, our own 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. At the highest level of our framework we assumed an understanding of entities, forces, and Newtonian-like dynamics. It is possible this level is either innate (Spelke & Kinzler, 2007) or extremely early developing, learned through processes outside those of the proposed framework.

Second, how does the language people use to talk about physical properties relate to quantitative descriptions of those properties? In our task and in day-to-day physical descriptions words like "heavy" or "rough" are used, which describe continuous qualities. These words are also graded adjectives with context-sensitive boundaries. An addition to our model could include drawing such properties from continuous distributions, such as different power-law distributions for the meaning of the words "light" and "heavy." We did not originally use such distributions because then even the ideal optimal inference model must be approximated rather than enumerated, as the space of continuous concepts cannot be searched and scored

[8] In order to facilitate the exploration of other features, the full participant responses as well as the trajectory data for all stimuli is available at https://tomerullman.org/physics-cogpsy-2017/data.html.

exhaustively. Such an approximation raises questions about the exact technique to use, without allowing us to compare between ideal and approximate techniques, but it is possible and worth exploring (see for example Ullman et al., 2012 on approximate search in large theory spaces).

Third, what kind of physical forces, properties and dynamics do people find natural? What is intuitive in intuitive physics? In our framework we considered the limited space of pairwise and global forces, friction, collisions, and stable conserved properties shared across objects, and people seemed able to reason about these relatively well. We believe people are able to reason about spring- and string-like forces, as well as attachments that maintain certain constraints on object relations. But it is entirely possible for the framework to generate and explore what we think will be non-intuitive dynamical scenes that people will find difficult to reason about, such as time-dependent, velocity-dependent forces that act according to non-conserved properties of objects. However, these forces would be more difficult to express in traditional physics simulations (in the sense that the code is longer to write), suggesting a possible link to explore between simplicity in description length and human reasoning in intuitive physics.

Finally, what are the perceptual inputs that go into physical reasoning? Are they simply pixels that get grouped into 'motion features' used for bottom-up classification, or are the inputs properties of objects? This question parallels the top-down vs. bottom-up debate of object recognition in visual perception, and like that debate it might turn out to not be an either-or distinction (see for example Lamme & Roelfsema, 2000; Ullman, 1995). For example, one can ask: how are useful motion features learned? The automatic discovery of useful features is a substantial problem in machine learning. Our framework suggests at least tentatively that new features for rapid classification might be partially discovered by using synthetic data, generated by running forward simulations from an intuitive physics model of the world, rather than relying on experience in the absence of such a model.

6. Conclusion

The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny…'
[- Isaac Asimov]

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 proposed a hierarchical Bayesian framework that represents intuitive physical theories in terms of probabilistic programs, to explain how these theories can be learned and used across different environments and multiple levels of abstraction. We examined learning at the lowest levels of this framework using a challenging task of jointly inferring several physical properties and laws from short dynamic scenes with multiple interactive objects. Although participants were far from ideal observers in our experiments, they were nonetheless able to make reasonable inferences about all aspects of a given scenario's physics, and these inferences could serve as important first steps guiding subsequent causal learning.

Much recent work on the development of intuitive theories has emphasized the crucial role that active interventions – and not only observational data – play in making causal learning possible. Likewise in science, experimental interventions – and not simply correlational studies – have long been the gold standard for testing causal hypotheses. Yet controlled experiments and other interventions are not the only mode by which scientists and children learn about the world. They may not even be the most important. As Asimov suggests, every truly novel discovery in science begins with a moment of observation, a 'That's funny…' moment, when a keen observer notices that something isn't quite as she expected, and differs from the usual course of events in a way that is not simply random but has some novel structure that calls for out exploration, experimentation, and ultimately explanation.

We believe that this is just as true in the development of intuitive theories as in the development of formal scientific theories, and our studies here have attempted to capture this first step of learning in the domain of intuitive dynamics. In our experiments, the 'That's funny…' moment might occur when two objects veer slightly off their straight-line course towards one another, or when an object slows down more than expected while moving over a colored surface. In our modeling, probabilistic programs express the knowledge by which people imagine how a scene might play out under different candidate physical laws or parameters, and how, if the scene departs from the imagined path, parts of the original program might be adjusted to account for the surprising data. These hypothetical adjustments become the hypotheses to be tested in subsequent experiments, and with luck, the seeds of "Eureka!"

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. We also thank three anonymous reviewers for helpful comments and suggestions.

Appendix A

The following figures show a static representation of all 60 scenarios used in the experiment, using 4 images per scenario as it unfolds over time (see Figs. 11 and 12).

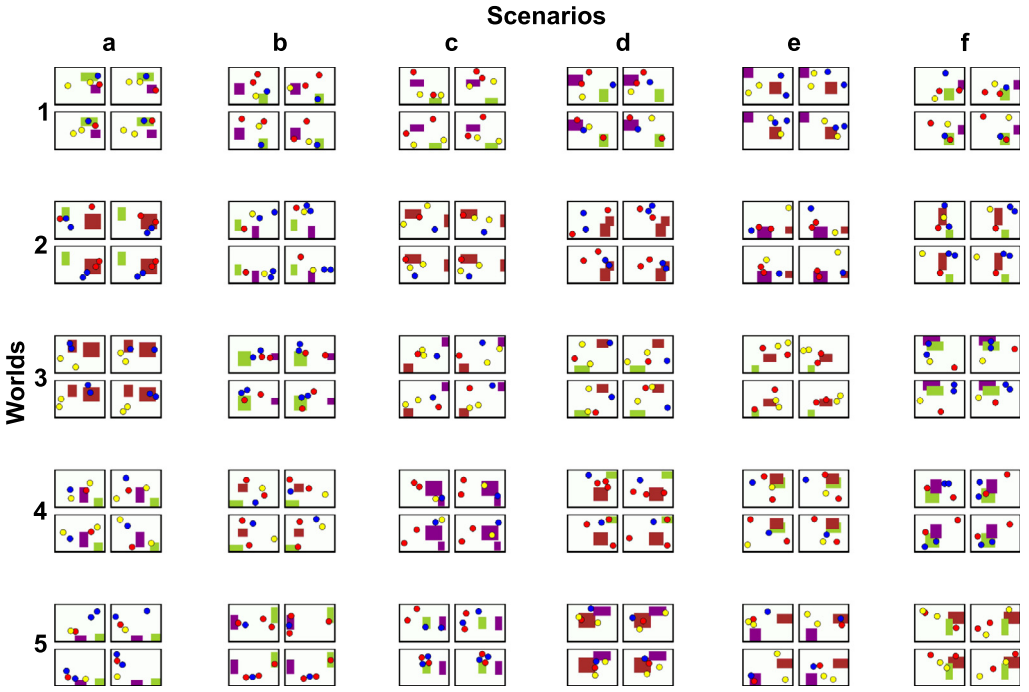

Fig. 11. Part 1 of all the stimuli used, showing 'worlds' 1–5 with 6 scenarios per world. There are 4 images per scenario, showing it unfold over time. The images were sampled at the start each scenario (upper left image in each scenario), 1.25 s into the scenario (upper right image), 3.75 s into the scenario (lower left image), and at the end of the scenario (5 s after it started, lower right image).

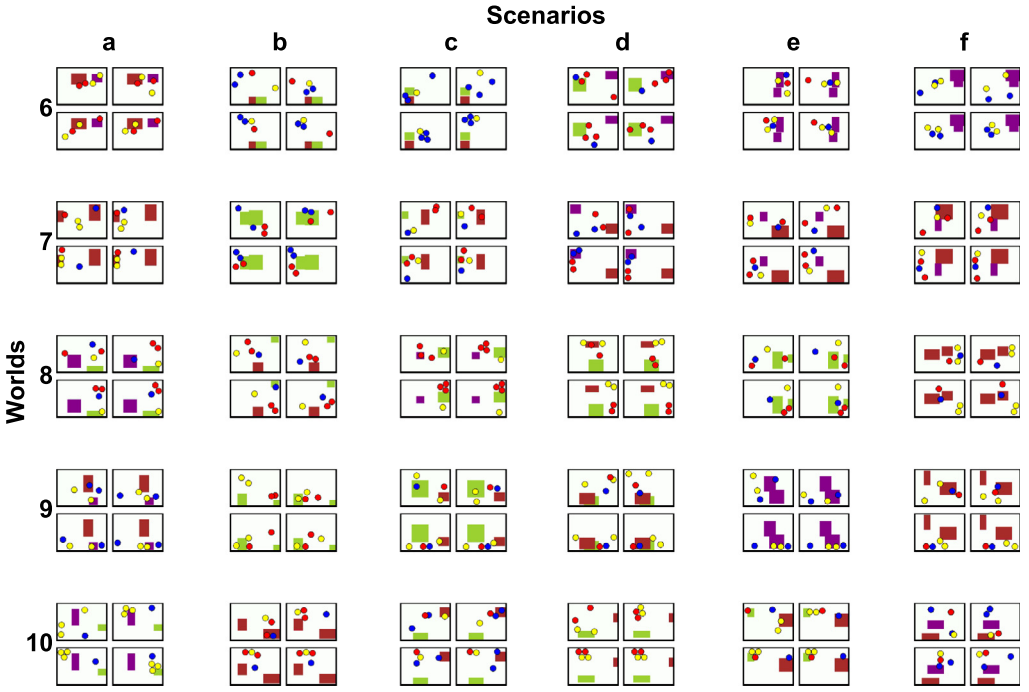

Fig. 12. Part 2 of all the stimuli used, showing 'worlds' 6–10 with 6 scenarios per world. There are 4 images per scenario, showing it unfold over time. The images were sampled at the start each scenario (upper left image in each scenario), 1.25 s into the scenario (upper right image), 3.75 s into the scenario (lower left image), and at the end of the scenario (5 s after it started, lower right image).

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*, 1–31.

Baillargeon, R. (2002). The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell Handbook of Childhood Cognitive Development*, 47–83.

Baillargeon, R. (2008). Innate ideas revisited: For a principle of persistence in infants' physical reasoning. *Perspectives on Psychological Science, 3*, 2–13.

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 of the United States of America, 110*, 18327–18332.

Battaglia, P., Pascanu, R., Lai, M., Jimenez Rezende, D., & Koray, K. (2016). Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems*. .

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

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in Cognitive Sciences, 18*, 497–500.

Carey, S. (2004). Bootstrapping and the origin of concepts. *Daedalus, 133*, 59–68.

Carroll, C. D., & Kemp, C. (2015). Evaluating the inverse reasoning account of object discovery. *Cognition, 139*, 130–153.

Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. (2017). A compositional object-based approach to learning physical dynamics. In *Proceedings of the 5th annual international conference on learning representations*. .

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54–75.

Forbus, K. D. (1988). Qualitative physics: Past, present, and future. In *Exploring artificial intelligence* (pp. 239–296).

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

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation, 24*, 1–24.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society*. .

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

Gilden, D. L., & Proffitt, D. R. (1994). Heuristic judgment of mass ratio in two-body collisions. *Perception & Psychophysics, 56*, 708–720.

Goodman, N. D., Frank, M. C., Griffiths, T. L., Tenenbaum, J. B., Battaglia, P., & Hamrick, J. (2015). Relevant and robust. A response to Marcus and Davis. *Psychological Science*.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In Margolis & Lawrence (Eds.), *The conceptual mind: New directions in the study of concepts*. MIT Press.

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*.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review, 118*, 110.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children.

Gopnik, A., & Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 17*, 1205–1222.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*, 1085.

Gourieroux, C. S., Monfort, A., & Renault, E. M. (1993). Indirect inference. *Journal of Applied Econometrics, 8*, S85–118.

Griffiths, T. L., Baraff, E. R., & Tenenbaum, J. B. (2004). Using physical theories to infer hidden causal structure. In *proceedings of the 26th annual conference of the cognitive science society* (pp. 500–505).

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21*, 263–268.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex physical scenes via probabilistic simulation. *Cognition, 1*, 2.

Hespos, S. J., Ferry, A. L., & Rips, L. J. (2009). Five-month-old infants have different expectations for solids and liquids. *Psychological Science, 20*, 603–611.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning, 37*, 183–233.

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

Kim, I. K., & Spelke, E. S. (1992). Infants' sensitivity to effects of gravity on visible object motion. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 385.

Kucukelbir, A., Ranganath, R., Gelman, A., & Blei, D. (2015). Automatic variational inference in Stan. In *Advances in neural information processing systems* (pp. 568–576).

Kulkarni, T., Yildirim, I., Kohli, P., Freiwald, W., & Tenenbaum, J. (2014). Deep generative vision as approximate Bayesian computation. In *NIPS 2014 ABC workshop*.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences, 23*, 571–579.

Lee, S. A., & Spelke, E. S. (2010). Two systems of spatial representation underlying navigation. *Experimental Brain Research, 206*, 179–188.

Lieder, F., Griffiths, T., Huys, Q. J., & Goodman, N. (2016). A rational perspective on anchoring-and-adjustment: The anchoring bias reflects rational use of cognitive resources. Unpublished Manuscript.

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In *Advances in neural information processing systems* (pp. 2690–2798).

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition, 131*, 284–299.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science, 24*, 2351–2360.

Marr, D., & Poggio, T. (1976). From understanding computation to understanding neural circuitry. *AI Memo, 357*, 1–22.

McIntyre, J., Zago, M., Berthoz, A., & Lacquaniti, F. (2001). Does the brain model Newton's laws? *Nature Neuroscience, 4*, 693–694.

Needham, A., & Baillargeon, R. (1993). Intuitions about support in 4.5-month-old infants. *Cognition, 47*, 121–148.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.

Rips, L. J., & Hespos, S. J. (2015). Divisions of the physical world: Concepts of objects and substances.

Ritchie, D., Horsfall, P., & Goodman, N. D. (2016). Deep amortized inference for probabilistic programs. Available from 1610.05735.

Runeson, S., Juslin, P., & Olsson, H. (2000). Visual perception of dynamic properties: Cue heuristics versus direct-perceptual competence. *Psychological Review, 107*, 525–555.

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

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

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences, 23*, 701–717.
Stickgold, R., Malia, A., Maguire, D., Roddenberry, D., & O'Connor, M. (2000). Replaying the game: Hypnagogic images in normals and amnesics. *Science (New York, NY), 290*, 350–353.
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 (New York, N.Y.), 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 (New York, N.Y.), 331*, 1279–1285.
Todd, J. T., & Warren, W. H. (1982). Visual perception of relative mass in dynamic events. *Perception, 11*, 325–335.
Ullman, S. (1995). Sequence seeking and counter streams: A computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex, 5*, 1–11.
Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of the thought. *Cognitive Development*.
Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences, 21*(9), 649–665.
Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science, 38*, 599–637.
Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*, 337–375.
Wingate, D., & Weber, T. (2013). Automated variational inference in probabilistic programming. Available from 1301.1299.
Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems* (pp. 127–135).