Simulation as an engine of physical scene understanding

Peter W. Battaglia[1], Jessica B. Hamrick, and Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved September 20, 2013 (received for review April 8, 2013)

In a glance, we can perceive whether a stack of dishes will topple, a branch will support a child's weight, a grocery bag is poorly packed and liable to tear or crush its contents, or a tool is firmly attached to a table or free to be lifted. Such rapid physical inferences are central to how people interact with the world and with each other, yet their computational underpinnings are poorly understood. We propose a model based on an "intuitive physics engine," a cognitive mechanism similar to computer engines that simulate rich physics in video games and graphics, but that uses approximate, probabilistic simulations to make robust and fast inferences in complex natural scenes where crucial information is unobserved. This single model fits data from five distinct psychophysical tasks, captures several illusions and biases, and explains core aspects of human mental models and common-sense reasoning that are instrumental to how humans understand their everyday world.

To see is, famously, "to know what is where by looking" (ref. 1, p. 3). However, to see is also to know what will happen and what can be done and to detect not only objects and their locations, but also their physical attributes, relationships, and affordances and their likely pasts and futures conditioned on how we might act. Consider how objects in a workshop scene (Fig. 1 *A* and *B*) support one another and how they respond to various applied forces. We see that the table supports the tools and other items on its top surface: If the table were removed, these objects would fall. If the table were lifted from one side, they would slide toward the other side and drop off. The table also supports a tire leaning against its leg, but precariously: If bumped slightly, the tire might fall. Objects hanging from hooks on the wall can pivot about these supports or be easily lifted off; in contrast, the hooks themselves are rigidly attached.

This physical scene understanding links perception with higher cognition: grounding abstract concepts in experience, talking about the world in language, realizing goals through actions, and detecting situations demanding special care (Fig. 1*C*). It is critical to the origins of intelligence: Researchers in developmental psychology, language, animal cognition, and artificial intelligence (2–6) consider the ability to intentionally manipulate physical systems, such as building a stable stack of blocks, as a most basic sign of humanlike common sense (Fig. 1*D*). It even gives rise to some of our most viscerally compelling games and art forms (Fig. 1 *E* and *F*).

Despite the centrality of these physical inferences, the computations underlying them in the mind and brain remain unknown. Early studies of intuitive physics focused on patterns of errors in explicit reasoning about simple one-body systems and were considered surprising because they suggested that human intuitions are fundamentally incompatible with Newtonian mechanics (7). Subsequent work (8, 9) has revised this interpretation, showing that when grounded in concrete dynamic perceptual and action contexts, people's physical intuitions are often very accurate by Newtonian standards, and pointing out that even in the earlier studies, the majority of subjects typically gave correct responses (10). Several recent models have argued that both successes and biases in people's perceptual judgments about simple one- and two-body interactions (e.g., judging the relative masses of two colliding point objects) can be explained as rational probabilistic inferences in a "noisy Newtonian" framework, assuming Newton's laws plus noisy observations (11–14). However, all of this work addresses only very simple, idealized cases, much closer to the examples of introductory physics classes than to the physical contexts people face in the real world. Our goal here is to develop and test a computational framework for intuitive physical inference appropriate for the challenges and affordances of everyday scene understanding: reasoning about large numbers of objects, only incompletely observed and interacting in complex, nonlinear ways, with an emphasis on coarse, approximate, short-term predictions about what will happen next.

Our approach is motivated by a proposal first articulated by Kenneth Craik (15), that the brain builds mental models that support inference by mental simulations analogous to how engineers use simulations for prediction and manipulation of complex physical systems (e.g., analyzing the stability and failure modes of a bridge design before construction). These runnable mental models have been invoked to explain aspects of high-level physical and mechanical reasoning (16, 17) and implemented computationally in classic artificial intelligence systems (18–20). However, these systems have not attempted to engage with physical scene understanding: Their focus on qualitative or propositional representations, rather than quantitative aspects and uncertainties of objects' geometry, motions, and force dynamics, is better suited to explaining high-level symbolic reasoning and problem solving. To understand physics in the context of scene perception and action, a more quantitative and probabilistic approach to formalizing mental models is required.

Here we introduce such a framework, which exploits recent advances in graphics and simulation tools, as well as Bayesian cognitive modeling (21), to explain how people understand the physical structure of real-world scenes. We posit that human judgments are driven by an "intuitive physics engine" (IPE), akin to the computer physics engines used for quantitative but approximate simulation of rigid body dynamics and collisions, soft body and fluid dynamics in computer graphics, and interactive video games. The IPE performs prediction by simulation and incorporates uncertainty about the scene by treating its simulation runs as statistical samples. We focus on how the IPE supports inferences about configurations of many rigid objects subject to gravity and friction, with varying numbers, sizes, and masses, like those typical in children's playrooms, office desktops, or the workshop, in Fig. 1*A*. In a series of experiments we show that the IPE can make numerous quantitative judgments that are surprisingly consistent with those of probabilistic physics simulations, but also that it differs from ground truth physics in crucial ways. These differences make the IPE more robust and useful in everyday cognition, but also prone to certain limitations and illusions (as in Fig. 1*F*).

Architecture of the IPE. We propose a candidate architecture for the IPE that can interface flexibly with both lower-level perceptuomotor systems and higher-level cognitive systems for

Author contributions: P.W.B., J.B.H., and J.B.T. designed research; P.W.B. and J.B.H. performed research; P.W.B. and J.B.H. analyzed data; and P.W.B. and J.B.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1] To whom correspondence should be addressed. E-mail: pbatt@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1306572110/-/DCSupplemental.

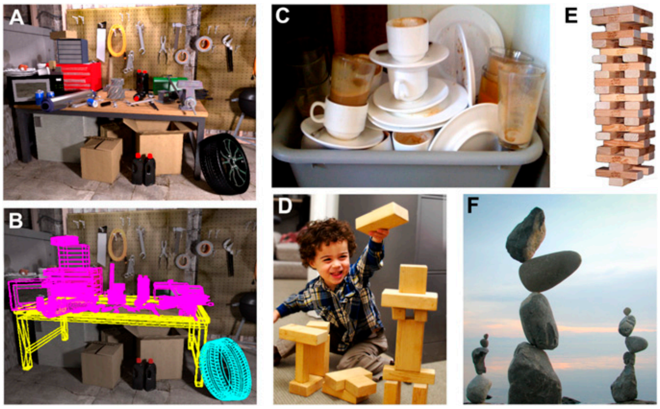

Fig. 1. Everyday scenes, activities, and art that evoke strong physical intuitions. (*A*) A cluttered workshop that exhibits many nuanced physical properties. (*B*) A 3D object-based representation of the scene in *A* that can support physical inferences based on simulation. (*C*) A precarious stack of dishes looks like an accident waiting to happen. (*D*) A child exercises his physical reasoning by stacking blocks. (*E*) Jenga puts players' physical intuitions to the test. (*F*) "Stone balancing" exploits our powerful physical expectations (Photo and stone balance by Heiko Brinkmann).

planning, action, reasoning, and language (Fig. 2*A*). At its core is an object-based representation of a 3D scene—analogous to the geometric models underlying computer-aided design programs (Fig. 1*B*)—and the physical forces governing the scene's dynamics: how its state changes over time (Fig. 2*A*). This representation quantitatively encodes a large number of static and dynamic variables needed to capture the motions and interactions of many objects. This may include objects' geometries, arrangements, masses, elasticities, rigidities, surface characteristics, and velocities, as well as the effects of forces acting on objects due to gravity, friction, collisions, and other potentially unobservable sources.

The IPE thus represents the world with a reasonable degree of physical fidelity. However, three key design elements render it distinct from an ideal physicist's approach and more akin to an engineer's. First, the IPE is based on simulation: Rather than manipulating symbolic equations to obtain analytic solutions, it represents mechanics procedurally and generates predicted states based on initial ones by recursively applying elementary physical rules over short time intervals. Second, the IPE is probabilistic rather than deterministic: It runs stochastic (Monte Carlo) simulations (22) that represent uncertainty about the scene's state and force dynamics and is thereby robust to the noisy and incomplete information provided by perception. Third, the IPE is inherently approximate: In its mechanics rules and representations of objects, forces, and probabilities, it trades precision and veridicality for speed, generality, and the ability to make predictions that are good enough for the purposes of everyday activities.

To make this proposal concrete and testable, we also need to specify the nature of these approximations and how coarse or fine grained they are. Here the IPE likely departs from engineering practice: People's everyday interactions with their surroundings often have much tighter time constraints and more relaxed fault tolerances, leading our brains to favor speed and generality over the degree of precision needed in engineering problems. Our initial IPE model thus adopts the simplest general-purpose approximation tools we know of. We used the Open Dynamics Engine (ODE) (www.ode.org) as a mechanism for approximate rigid-body dynamics simulations and the most naive Monte Carlo approach of black-box forward simulation (22) as a mechanism for representing and propagating approximate probabilities through these physical dynamics. The ODE represents objects' geometries as polyhedra and their mass distributions by inertial tensors, and its simulations do not enforce the conservation of energy or momentum explicitly, but only implicitly via coarse event detection and resolution procedures. Our model runs the simulator on multiple independent draws from the observer's probability distribution over scenes and forces to form an approximate posterior distribution over future states over time. Even within the range of speed–accuracy tradeoffs that our initial IPE model supports, we expect that people will tend to adopt the cheapest approximations possible (see *SI Appendix: Approximations*). The IPE may dramatically simplify objects' geometry, mass density distributions, and physical interactions, relative to what the ODE allows; and instead of running many Monte Carlo simulations, the IPE may encode probabilities very coarsely by using only one or a few samples (as people do in simpler decision settings) (23).

Our central claim is that approximate probabilistic simulation plays a key role in the human capacity for physical scene understanding and can distinctively explain how people make rich inferences in a diverse range of everyday settings, including many that have not previously been formally studied. Given an appropriate geometric model (Fig. 1*B*) of the workshop scene in Fig. 1*A*, the IPE can compute versions of many of the intuitive inferences about that scene described above. Given a geometric model of the scene in Fig. 1*C*, it can explain not only how we infer that the stacked dishes are precarious, but also how we can answer many other queries: Which objects would fall first? How might they fall—in which direction, or how far? Which other objects might they cause to fall? Everyday scenarios can exhibit great variety in objects' properties (e.g., their weight, shape, friction, etc.) and the extrinsic forces that could be applied (e.g., from a slight bump to a jarring blow), and our IPE model can capture how people's predictions are sensitive to these factors—including ways that go beyond familiar experience. In Fig. 1*C*, for instance, we can infer that a cast-iron skillet placed onto the dishes would be far more

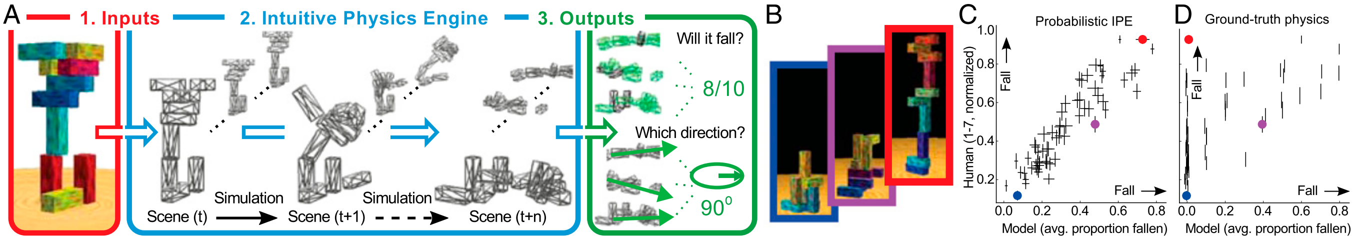

Fig. 2. (*A*) The IPE model takes inputs (e.g., perception, language, memory, imagery, etc.) that instantiate a distribution over scenes (*1*), then simulates the effects of physics on the distribution (*2*), and then aggregates the results for output to other sensorimotor and cognitive faculties (*3*). (*B*) Exp. 1 (Will it fall?) tower stimuli. The tower with the red border is actually delicately balanced, and the other two are the same height, but the blue-bordered one is judged much less likely to fall by the model and people. (*C*) Probabilistic IPE model (*x* axis) vs. human judgment averages (*y* axis) in Exp. 1. See Fig. S3 for correlations for other values of σ and ϕ. Each point represents one tower (with SEM), and the three colored circles correspond to the three towers in *B*. (*D*) Ground truth (nonprobabilistic) vs. human judgments (Exp. 1). Because it does not represent uncertainty, it cannot capture people's judgments for a number of our stimuli, such as the red-bordered tower in *B*. (Note that these cases may be rare in natural scenes, where configurations tend to be more clearly stable or unstable and the IPE would be expected to correlate better with ground truth than it does on our stimuli.)

destabilizing than a paper plate or that placing these stacked dishes near the edge of a table would be much less wise if there were children running about than if the room were empty. Such intuitions come naturally and (fortunately) do not require that we experience each of these situations firsthand to be able to understand them. Together, these types of inferences constitute an answer to the more general question, “What will happen?”, that humans can answer across countless scenes and that can be read off from the IPE’s simulations.

Psychophysical Experiments. Relative to most previous research on intuitive physics, our experiments were designed to be more representative of everyday physical scene understanding challenges, similar to those shown in Fig. 1 and discussed above. These tasks feature complex configurations of objects and require multiple kinds of judgments in different output modalities and graded (rather than simply all-or-none, yes-or-no) predictions, yet are still constrained enough to allow for controlled quantitative psychophysical study. Our most basic task (Exp. 1) probed people’s judgments of stability by presenting them with towers of 10 blocks arranged in randomly stacked configurations (Fig. 2*B*) and asking them to judge (on a 1–7 scale) “Will this tower fall?” under the influence of gravity. After responding, observers received visual feedback showing the effect of gravity on the tower, i.e., whether and how the blocks of the tower would fall under a ground truth physics simulation.

The critical test of our IPE account is not whether it can explain every detail of how people respond in one such task, but whether it can quantitatively explain the richness of people’s intuitions about what will happen across a diverse range of tasks. Hence subsequent experiments manipulated elements of Exp. 1 to examine whether the model could account for people’s ability to make different predictions about a given scene (Exps. 2 and 4), their sensitivity to underlying physical attributes such as mass (Exps. 3 and 4), and their ability to generalize to a much wider and more complex range of scenes (Exp. 5).

Applying our IPE model to these tasks requires choices about how to formalize each task’s inputs and outputs—how each stimulus gives rise to a sample of initial object states and force dynamics for the simulator and how the effects of simulated physics on this sample are used to make the task’s judgment—as well as choices about the specifics of the simulation runs. Although the “Will it fall?” task primarily involved visual inputs and linguistic outputs, later tasks (Exps. 2–5) examined the flexibility of the IPE’s interfaces with other cognitive systems by adding linguistic inputs, symbolic visual cues, and sensorimotor outputs. To allow the same core IPE model to be testable across all experiments, we made the following simplifying assumptions to summarize these other interfaces.

We set the IPE’s input to be a sample from a distribution over scene configurations, object properties, and forces based on ground truth, but modulated by a small set of numerical parameters that capture ways in which these inputs are not fully observable and might vary as a function of task instructions. The first parameter, σ, captures uncertainty in the observer’s representation of the scene’s initial geometry—roughly, as the SD of a Bayesian observer’s posterior distribution for each object’s location in 3D space, conditioned on the 2D stimulus images. The second parameter, ϕ, reflects the magnitude of possible latent forces that the observer considers could be applied (e.g., a breeze, a vibration, or a bump) to the objects in the scene, in addition to those forces always known to be present (e.g., gravity, friction, and collision impacts). The third parameter, μ, captures physical properties that vary across objects but are not directly observable—specifically, the relative mass of different objects—but other properties such as elasticity or surface roughness could be included as well.

Given such an input sample, our IPE model simulated physical dynamics to produce a sample of final scene configurations. In some cases the objects moved due to gravitational or external forces or ensuing secondary collisions, whereas in others they remained at their initial state. The model’s output consists of aggregates of simple spatial, numerical, or logical predicates applied to the simulation runs, as appropriate for the task and judgment (*SI Appendix: IPE Model*). For example, for the Will it fall? query, we took the IPE’s output to be the average proportion of blocks that fell across the simulation runs.

Each manipulation in Exps. 1–5 tested the IPE model in increasingly complex scenarios, which the model accommodates by adjusting its manipulation-sensitive input parameters or output predicates; all manipulation-irrelevant model components are fixed to previously fitted values. We also contrasted the model with variants insensitive to these manipulations, to assess how fully the IPE represents these physical, scene, and task features. Finally, we explored several ways in which the human IPE might adopt even simpler approximate representations.

Results

Exp. 1: Will It Fall? Exp. 1 measured each subject’s ($n = 13$) Will it fall? judgments about 60 different tower scenes, repeated six times over separate blocks of trials (see *SI Materials and Methods*, Fig. S1, and Table S1). Fig. 2*C* shows the correlation between the model’s and people’s average judgments ($\rho = 0.92[0.88, 0.94]$, where $[l, u]$ indicates lower/upper 95% confidence intervals) under the best-fit input parameters: $\sigma = 0.2$, or 20% of the length of a block’s shorter side, and $\phi = 0.2$, corresponding to very small applied external forces, on the scale of a light tap. Nearby values of σ and ϕ also had high correlations because state and force uncertainty influenced the model’s predictions in similar ways (Fig. S3). The μ parameter was set to 1 because all objects had identical physical properties. We analyzed subjects’ responses for improvements across trial blocks and found no effects of either the amount of feedback or the amount of practice (Fig. S7 and *SI Appendix: Analysis of Learning*). We also replicated the design of Exp. 1 on a new group of subjects ($n = 10$) who received no feedback and found their mean responses to be highly correlated with those in the original feedback condition ($\rho = 0.95[0.95, 0.95]$), confirming that any feedback-driven learning played at most a minimal role.

To assess the role of probability in the IPE simulations, we also compared people’s judgments to a deterministic ground truth physics model (the same simulations that were used to provide posttrial feedback). This ground truth model corresponds to a variant of the IPE model where $\sigma = 0$ and $\phi = 0$ (i.e., each simulation is run with initial states identical to the true objects’ states and uses no forces besides gravity, friction, and collisions). The task was challenging for subjects: Their average accuracy was 66% (i.e., percentage of their thresholded responses matching the ground truth model), and their correlation with the ground truth predictions was significantly lower ($\rho = 0.64[0.46, 0.79]$, $P < 0.001$; Fig. 2*D*) than with the IPE model. This demonstrates the crucial role of including state and force uncertainty in the model’s simulations and explains illusions like the surprisingly balanced stones in Fig. 1*F*: The ground truth scene configuration is in fact balanced, but so delicately that most similar configurations (and hence most of the IPE’s probabilistic simulations) are unbalanced and fall under gravity. We included an analogous illusory stimulus in the experiment, a delicately balanced tower (Fig. 2*B*, red border) that in fact stands up under ground truth physics but that the IPE model’s probabilistic simulations predict is almost certain to fall. As predicted by the IPE model, but not the ground truth variant, people judged this to be one of the most unstable towers in the entire stimulus set (Fig. 2 *C* and *D*, red circle).

Is it possible that people’s judgments did not involve any mental simulation at all, probabilistic or otherwise? We also tested an alternative account in the spirit of exemplar-based models and simple heuristics that have been proposed in previous studies of physical judgments (8–11): that people might instead base their judgments exclusively on learned combinations of geometric features of the initial scene configuration (e.g., the numbers, positions, and heights of the objects; see Table S2) without explicit reference to physical dynamics. This “feature-based” account consistently fared worse at predicting people’s judgments than the IPE model—sometimes dramatically worse (Fig. S4)—in Exp. 1

and a controlled follow-up experiment (Exp. S1) (*SI Appendix: Model-Free Accounts*) in which the towers were all of the same height, as well as in Exps. 2–5 described below. This is not to claim that geometric features play no role in physical scene understanding; in *SI Appendix: Approximations*, we describe settings where they might. However, our results show that they are not viable as a general-purpose alternative to the IPE model.

Exp. 2: In Which Direction? To test the IPE model's ability to explain different judgments in different modalities, we showed subjects ($n = 10$) scenes similar to those in Exp. 1, but instead asked them to judge the direction in which the tower would fall (Fig. 3*A* and Fig. S2). The IPE model's output predicate for this "In which direction?" query was defined as the angle of the average final position of the fallen blocks; input parameters ($\sigma = 0.2$, $\phi = 0.2$) and all other details were set to those used in modeling Exp. 1. Model predictions were very accurate overall: Subjects' mean direction judgments were within ±45° of the model's for 89% of the tower stimuli (Fig. 3*B*). As in Exp. 1, capturing uncertainty was crucial: The circular correlation with people's judgments was significantly higher for the IPE model ($\rho_{circ} = 0.80[0.71, 0.87]$) than for the ground-truth ($\sigma = 0$, $\phi = 0$) model (Fig. 3*C*; $\rho_{circ} = 0.61[0.46, 0.75]$, $P < 0.001$). These results show how a single set of probabilistic simulations from the IPE can account for qualitatively different types of judgments about a scene simply by applying the appropriate output predicates.

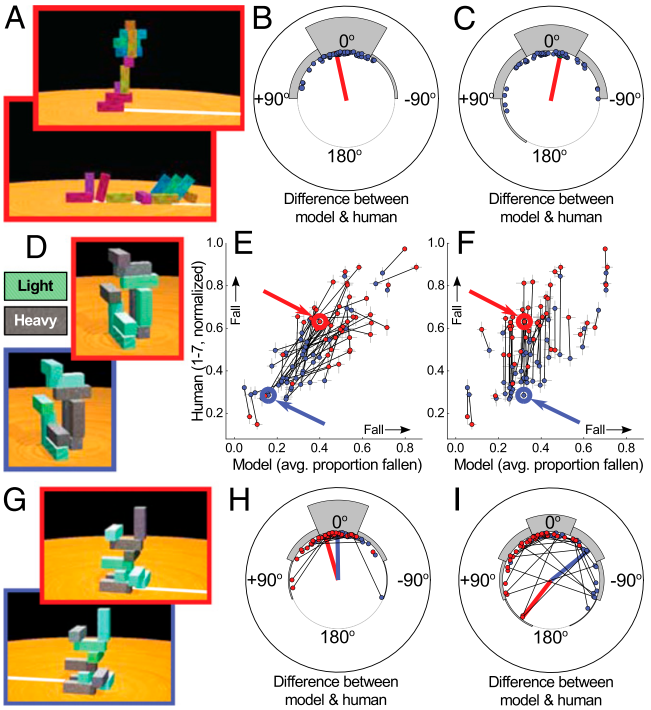

Fig. 3. (*A*) Exp. 2 (In which direction?). Subjects viewed the tower (*Upper*), predicted the direction in which it would fall by adjusting the white line with the mouse, and received feedback (*Lower*). (*B*) Exp. 2: Angular differences between the probabilistic IPE model's and subjects' circular mean judgments for each tower (blue points), where 0 indicates a perfect match. The gray bars are circular histograms of the differences. The red line indicates the tower in *A*. (*C*) The same as *B*, but for the ground truth model. (*D*) Exp. 3 (Will it fall?: mass): State pair stimuli (main text). Light blocks are green, and heavy ones are dark. (*E*) Exp. 3: The mass-sensitive IPE model's vs. people's judgments, as in Fig. 2*C*. The black lines connect state pairs. Both model and people vary their judgments similarly within each state pair (lines' slopes near 1). (*F*) Exp. 4: The mass-insensitive model vs. people. Here the model cannot vary its judgments within state pairs (lines are near vertical). (*G*) Exp. 4 (In which direction?: mass): State pair stimuli. (*H*) Exp. 4: The mass-sensitive IPE model's vs. people's judgments, as in *B*. The black lines connect state pairs. The model's and people's judgments are closely matched within state pairs (short black lines). (*I*) Exp. 4: The mass-insensitive IPE model vs. people. Here again, the model cannot vary its judgments per state pair (longer black lines).

Exps. 3 and 4: Varying Object Masses. To test the sensitivity of people's predictions to objects' physical attributes and the IPE model's ability to explain this sensitivity, Exps. 3 and 4 used designs similar to Exps. 1 and 2, respectively, but with blocks that were either heavy or light (10:1 mass ratio, indicated visually by different block colors; Fig. 3 *D* and *G*). We created pairs of stimuli ("state pairs") that shared identical geometric configurations, but that differed by which blocks were assigned to be heavy and light (Fig. 3 *D* and *G*) and thus in whether, and how, the blocks should be expected to fall. Again the IPE model's input parameters and output predicates were set identically to those used in Exps. 1 and 2, except that the mass parameter, μ, could vary to reflect people's understanding of the ratio between heavy and light blocks' masses. At the best-fitting value from Exp. 3, $\mu = 8$, model fits for Exp. 3 (Will it fall? judgment; Fig. 3*E*, $\rho = 0.80[0.72, 0.86]$) and Exp. 4 (In which direction? judgment; Fig. 3*H*, $\rho_{circ} = 0.78[0.67, 0.87]$) were comparable to those in Exps. 1 and 2, respectively; the true mass ratio ($\mu = 10$) yielded almost identical predictions and fits. By contrast, using the mass-insensitive ($\mu = 1$) model variant yielded significantly worse fits for both Exp. 3 (Fig. 3*F*, $\rho = 0.63[0.50, 0.73]$, $P < 0.001$) and Exp. 4 (Fig. 3*I*, $\rho_{circ} = 0.41[0.27, 0.57]$, $P < 0.001$). Differences in judgments about towers within each state pair also covaried significantly for people and the IPE model in both experiments (Exp. 3, $\rho = 0.73[0.62, 0.81]$; Exp. 4, $\rho_{circ} = 0.50[0.18, 0.75]$), whereas for the mass-insensitive model variants these correlations were 0 by definition. Together, these results show that people can incorporate into their predictions a key latent physical property that varies across objects (and is indicated only by covariation with a superficial color cue), that they do so in a near-optimal manner, and that the same IPE model could exploit the richer aspects of its scene representations to explain these inferences at a similar level of quantitative accuracy to that for the simpler tasks of Exps. 1 and 2 in which all objects were identical.

Exp. 5: Varying Object Shapes, Physical Obstacles, and Applied Forces. Exp. 5 was designed to be a comprehensive and severe test of the IPE model, evaluating how well it could explain people's judgments on a much more novel task in much more complex and variable settings—scenes with different sizes, shapes, numbers, and configurations of objects, with variable physical constraints on objects' motion due to attached obstacles and with added uncertainty about the external forces that could perturb the scene. Each scene depicted a table on which a collection of blocks were arranged (Fig. 4 *A* and *B*), half of which were red and the other half of which were yellow. Subjects ($n = 10$) were asked to imagine that the table is bumped hard enough to knock one or more of the blocks onto the floor and to judge which color of blocks would be more likely to fall off, using a 1–7 scale of confidence spanning "definitely yellow" to "definitely red". The 60 different scenes were generated by crossing 12 different block configurations—varying the numbers and shapes of the blocks and the numbers, heights, and positions of the stacks in which they were arranged—with five different tables, one with a flat surface and four others each with two short obstacles rigidly attached to different edges that interacted with the objects' motions in different ways (Fig. 4*A*). Two conditions differed in what information subjects received about the external bump: In the "cued" condition, a blue arrow indicated a specific direction for which subjects should imagine a bump; in the "uncued" condition, no arrow was shown and subjects had to imagine the effects of a bump from any possible direction (Fig. 4*B*). In the cued condition, each scene was shown with two different bump cue directions ("cue-wise pairs"). In 10 initial trials, subjects were familiarized with the task and the effects of a random bump strong enough to knock off at least one block, using simpler scenes for which the red–yellow judgment was obvious and the effect of the bump (applied for 200 ms) was shown after each judgment. Analogous feedback was also shown after every fifth experimental trial.

The IPE model was identical to that in Exps. 1 and 2 ($\sigma = 0.2, \mu = 1$), except for two differences appropriate for this task. To incorporate instructions about how the table is bumped, the magnitude of imagined external forces ϕ was increased to a range of values characteristic of the bumps shown during the familiarization period. The model simulated external forces under a range of magnitudes, varying in their effects from causing only a few blocks to fall off the table to causing most to fall off. For the uncued condition the model simulated all bump directions, whereas for the cued condition it simulated only bumps with directions within 45° of the cued angle (Fig. 4 *C* and *D*). The model's output predicate was defined as the proportion of red vs. total blocks that fell off the table, averaged across simulations.

Model predictions were strongly correlated with people's judgments in both the uncued and the cued bump conditions (Fig. 4*E*, $\rho = 0.89[0.82, 0.93]$, and Fig. 4*G*, $\rho = 0.86[0.80, 0.90]$, respectively). Fits were both qualitatively and quantitatively better than for model variants that did not take into account the obstacles (Figs. 4*F*, $\rho = 0.68[0.51, 0.81]$, $P < 0.002$; Fig. 4*H*, $\rho = 0.64[0.47, 0.77]$, $P < 0.001$), the bump cues (Fig. 4*I*, $\rho = 0.82[0.75, 0.87]$, $P < 0.2$), or either factor (Fig. 4*J*, $\rho = 0.58[0.41, 0.72]$, $P < 0.001$), suggesting both factors played causal roles in the IPE model's success. The model could also predict the effects of different obstacles and bump cues on people's judgments, with correlations of $\rho = 0.88[0.81, 0.93]$ between people's and the model's obstacle-wise differences in the uncued condition and $\rho = 0.64[0.46, 0.77]$ between their cue-wise differences in the cued condition. That the IPE model predicted judgments for these variable and complex scenarios at such high levels, comparable to the simpler experiments above, provides the strongest evidence yet that our model captures people's capacity for rich mental simulations of the physical world.

Approximations. Whereas the IPE model tested above attempts to represent scene structure, physical dynamics, and probabilities faithfully, given the constraints of a simple simulation engine and Monte Carlo inference scheme, the human IPE is likely bounded by further resource constraints and may adopt even coarser approximations. For example, instead of using many simulation samples to represent a full posterior predictive distribution, people might base their predictions on only very few samples. We estimated the number of samples that contribute to a subject's judgment by comparing the variance in subjects' responses to the variance in the model's responses, under the assumption that as the IPE pools more samples its trial-by-trial variance will decrease, and found that people's judgments were consistent with having been based on roughly three to seven stochastic simulation samples (*SI Appendix: Approximating Probabilities* and Fig. S6 *A–E*). We also compared IPE model variants that were limited to these small sample sizes to the large-sample models tested above and found that even these small sample sizes were sufficient to approximate well the predictive probability distributions in our tasks (Fig. S6 *F–J*). In other analyses, we found that people may fall back on non-simulation–based heuristics when simulations would require too much time and precision to be useful (*SI Appendix: Approximating Physics*) and that biases in how people predict the motions of nonconvex objects (10, 24) can be explained by an IPE that estimates objects' unknown mass distributions cheaply, using simplified geometric priors. Although preliminary, these results suggest that across a range of scenes and tasks, even a small number of coarse probabilistic simulations over short time intervals can support effective physical inferences and predict well people's judgments.

Discussion

We proposed that people's physical scene understanding can be explained by a simulation-based IPE that we formalized and tested in a wide range of experiments. This IPE model accounted well for diverse physical judgments in complex, novel scenes, even in the presence of varying object properties such as mass and uncertain external forces that could perturb the scene. Variants of the IPE model that were not sensitive to these physical differences consistently fit less well, as did combinations of special-purpose geometric features that did not model physics and had to be tailored to each experiment (Fig. S4 and *SI Appendix: Model-Free Accounts*), further supporting the case that human intuitions are driven by rich

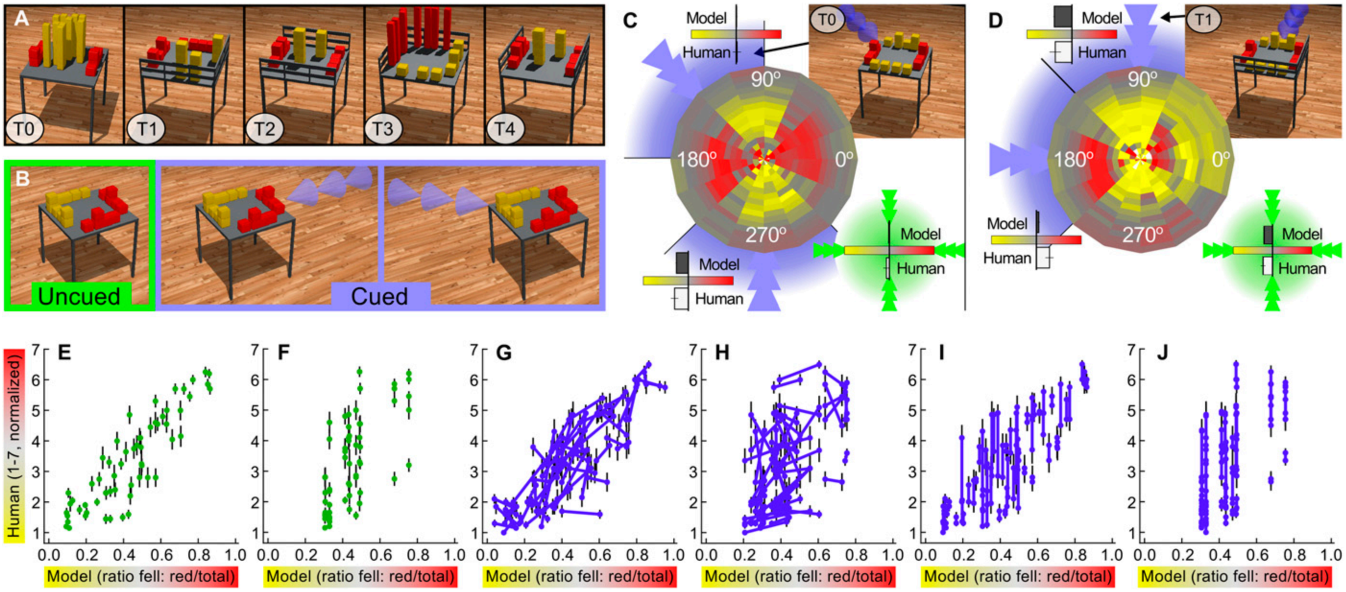

Fig. 4. Exp. 5 (Bump?). (*A*) Scene stimuli, whose tables have different obstacles (T0–T4). (*B*) In the uncued bump condition, subjects were not informed about the direction from which the bump would strike the scene; in the cued bump conditions, a blue arrowhead indicated the bump's direction. (*C*) The disk plot shows IPE model predictions per bump direction (angle) and ϕ (radius) for the stimulus in the image; the blue arrowheads/arcs indicate the range of bump angles simulated per bump cue, and the green circle and arrowheads represent the uncued condition. *Inset* bar graphs show the model's and people's responses, per cue/condition. (*D*) The same block configuration as in *C*, with different obstacles (T1). (*E–J*) IPE model's (*x* axis) vs. people's (*y* axis) mean judgments (each point is one scene, with SEM). The lines in *G–J* indicate cue-wise pairs. Each subplot show one cue condition and IPE model variant (correlations in parentheses, with *P* value of difference from full IPE): (*E*) Uncued, full IPE. (*F*) Uncued, obstacle insensitive (model assumes T0). (*G*) Cued, full IPE. (*H*) Cued, obstacle insensitive. (*I*) Cued, cue insensitive (model averages over all bump angles). (*J*) Cued, obstacle and cue insensitive.

physical simulations. That these simulations are probabilistic was strongly supported by the systematic deviations of people's judgments from ground truth physical simulations (the $\sigma = 0, \phi = 0$ model), as well as the existence of certain stability illusions (Fig. 1*F* and Fig. 2 *B–D*), all of which are naturally explained by the incorporation of uncertainty. Other illusions and patterns of error (Exp. S2 and Fig. S5) point to other ways in which these simulations approximate physical reality only coarsely, yet effectively enough for most everyday action-planning purposes. Probabilistic approximate simulation thus offers a powerful quantitative model of how people understand the everyday physical world.

This proposal is broadly consistent with other recent proposals that intuitive physical judgments can be viewed as a form of probabilistic inference over the principles of Newtonian mechanics (the noisy Newton hypothesis) (11–14). Previous noisy Newton models have been restricted to describing few judgments in simple scenarios (e.g., one or two point-like objects moving in one or two dimensions). Our work differs primarily in its focus on simulation—specifically rich, 3D, object-based simulations—as the means by which physical knowledge is represented and probabilistic inference is carried out. Our model can describe numerous judgments about complex natural scenes, both familiar and novel, and offers a plausible algorithmic basis for how people can make these judgments.

How else might people's physical scene understanding work, if not through model-based simulation? Much recent research in computer vision is based on model-free, data-driven approaches, which depend heavily on learning from past experience, either by memorizing very large labeled sets of exemplars or by training combinations of compact image features to predict judgments of interest. We do not argue against a role for memory or learned features in physical scene understanding, yet our results suggest that combinations of the most salient features in our scenes are insufficient to capture people's judgments (*SI Appendix: Model-Free Accounts* and Fig. S4). More generally, a purely model-free account seems implausible on several grounds: It would have to be flexible enough to handle a wide range of real-world scenes and inferences, yet compact enough to be learnable from people's finite experience. It would also require additional control mechanisms to decide which features and judgment strategies are appropriate for each distinct context, and it would be challenged to explain how people perform novel tasks in unfamiliar scenes or how their physical understanding might interface with their rich language, reasoning, imagination, and planning faculties. In contrast, model-based reasoning is more flexible and general purpose and does not require substantial task-specific learning. We know of no other approach that is a plausible competitor for making physical inferences and predicting What will happen? in everyday scenarios—let alone one that can quantitatively match the IPE model's consistency with people's judgments across our range of experiments. However, we encourage alternatives that can compete with our account and have made our stimuli and data freely available online for that purpose.

The generality of a simulation-based IPE goes well beyond the settings studied here. A more realistic visual front end can be added to capture people's perceptual uncertainty (due to viewpoint, lighting, or image occlusions; *SI Appendix: Bayesian Vision System* and Fig. S8) and working memory and attentional constraints (25). In ongoing work we are finding that the same IPE model can explain how people learn about the latent properties of objects (e.g., mass and friction) from observing their dynamics, how people infer attachment relations among objects in a scene, and how people plan actions to achieve desired physical outcomes. Its underlying knowledge of physics can also be extended to make inferences about the dynamics of other entity types (nonrigid objects, nonsolid substances, and fluids) that are not handled by the ODE, but can be instantiated in more sophisticated simulation engines such as Bullet or Blender.

More broadly, our work opens up unique directions for connecting people's understanding of physical scenes with other aspects of cognition. Probabilistic simulations may help explain how physical knowledge influences perceived scene layouts (26–28), movement planning (29), causal inferences (11, 12), language semantics, and syntax (e.g., "force dynamics") (4) and infants' expectations about objects (2, 30). Most generally, probabilistic simulation offers a way to integrate symbolic reasoning and statistical inference—two classically competing approaches to formalizing common-sense thought. The result is a framework that is both more quantitative and more amenable to rigorous psychophysical experimentation than previous accounts of human mental models and also better able to explain how people apprehend and interact with the physical environments they inhabit.

ACKNOWLEDGMENTS. We thank N. Kanwisher, E. Vul, T. Ullman, C. Baker, E. Davis, G. Marcus, and J. M. Tenenbaum for suggestions on the manuscript and overall research. Research support was provided by the National Institutes of Health 5F32EY019228-02; by Office of Naval Research Grants N00014-09-0124, N00014-07-1-0937, and 1015GNA126; by Qualcomm; and by Intelligence Advanced Research Project Activity Grant D10PC20023.

1. Marr D (1982) *Vision* (Freeman, San Francisco).
2. Baillargeon R (2002) The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell Handbook of Childhood Cognitive Development* (Blackwell, Oxford), Vol 1, pp 46–83.
3. Spelke ES, Breinlinger K, Macomber J, Jacobson K (1992) Origins of knowledge. *Psychol Rev* 99(4):605–632.
4. Talmy L (1988) Force dynamics in language and cognition. *Cogn Sci* 12(1):49–100.
5. Tomasello M (1999) *The Cultural Origins of Human Cognition* (Harvard Univ Press, Cambridge, MA).
6. Winston PH (1975) *The Psychology of Computer Vision* (McGraw-Hill, New York), Vol 73.
7. McCloskey M, Caramazza A, Green B (1980) Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science* 210(4474):1139–1141.
8. Gilden DL, Proffitt DR (1989) Understanding collision dynamics. *J Exp Psychol Hum Percept Perform* 15(2):372–383.
9. Nusseck M, Lagarde J, Bardy B, Fleming R, Bülthoff H (2007) Perception and prediction of simple object interactions. *Proceedings of the ACM Symposium on Applied Perception*, eds Wallraven C, Sundstedt V (Association for Computing Machinery, New York), pp 27–34.
10. Proffitt DR, Kaiser MK, Whelan SM (1990) Understanding wheel dynamics. *Cognit Psychol* 22(3):342–373.
11. Sanborn AN, Mansinghka VK, Griffiths TL (2013) Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychol Rev* 120(2):411–437.
12. Gerstenberg T, Goodman N, Lagnado D, Tenenbaum J (2012) Noisy newtons: Unifying process and dependency accounts of causal attribution. *Proceedings of the 34th Conference of the Cognitive Science Society*, eds Miyake N, Peebles D, Cooper RP (Cognitive Science Society, Austin, TX), pp 378–383.
13. Smith KA, Vul E (2013) Sources of uncertainty in intuitive physics. *Top Cogn Sci* 5(1):185–199.
14. Smith K, Battaglia P, Vul E (2013) Consistent physics underlying ballistic motion prediction. *Proceedings of the 35th Conference of the Cognitive Science Society*, eds Knauff M, Pauen M, Sebanz N, Wachsmuth I (Cognitive Science Society, Austin, TX), pp 3426–3431.
15. Craik K (1943) *The Nature of Explanation* (Cambridge Univ Press, Cambridge, UK).
16. Gentner D, Stevens A (1983) *Mental Models* (Lawrence Erlbaum, Hillsdale, NJ).
17. Hegarty M (2004) Mechanical reasoning by mental simulation. *Trends Cogn Sci* 8(6):280–285.
18. Johnson-Laird P (1983) *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Cambridge Univ Press, Cambridge, UK), Vol 6.
19. De Kleer J, Brown J (1984) A qualitative physics based on confluences. *Artif Intell* 24(1):7–83.
20. Forbus K (2011) Qualitative modeling. *Wiley Interdiscip Rev Cogn Sci* 2:374–391.
21. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011) How to grow a mind: Statistics, structure, and abstraction. *Science* 331(6022):1279–1285.
22. Rubinstein RY (2009) *Simulation and the Monte Carlo Method* (Wiley, New York), Vol 190.
23. Vul E, Goodman N, Griffiths T, Tenenbaum J (2009) One and done? Optimal decisions from very few samples. *Proceedings of the 31st Conference of the Cognitive Science Society*, eds Taatgen N, van Rijn H (Cognitive Science Society, Austin, TX), pp 66–72.
24. Cholewiak SA, Fleming RW, Singh M (2013) Visual perception of the physical stability of asymmetric three-dimensional objects. *J Vis* 13(4):12.
25. Vul E, Frank M, Alvarez G, Tenenbaum J (2009) Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Adv NIPS* 22:1955–1963.
26. Biederman I, Mezzanotte RJ, Rabinowitz JC (1982) Scene perception: Detecting and judging objects undergoing relational violations. *Cognit Psychol* 14(2):143–177.
27. Freyd JJ (1987) Dynamic mental representations. *Psychol Rev* 94(4):427–438.
28. Hock H, Gordon G, Whitehurst R (1974) Contextual relations: The influence of familiarity, physical plausibility, and belongingness. *Atten Percept Psychophys* 16(1):4–8.
29. Zago M, McIntyre J, Senot P, Lacquaniti F (2009) Visuo-motor coordination and internal models for object interception. *Exp Brain Res* 192(4):571–604.
30. Téglás E, et al. (2011) Pure reasoning in 12-month-old infants as probabilistic inference. *Science* 332(6033):1054–1059.