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.

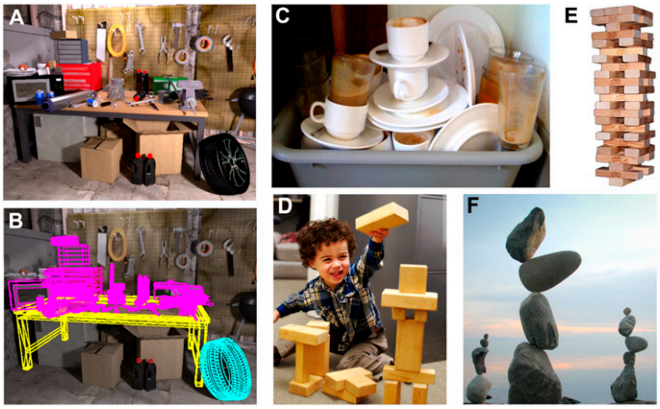

Fig. 1. Everyday scenes, activities, and art that evoke strong physical intuitions. (*A*) A cluttered workshop that exhibits many nuanced physical properties. (*B*) A 3D object-based representation of the scene in *A* that can support physical inferences based on simulation. (*C*) A precarious stack 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

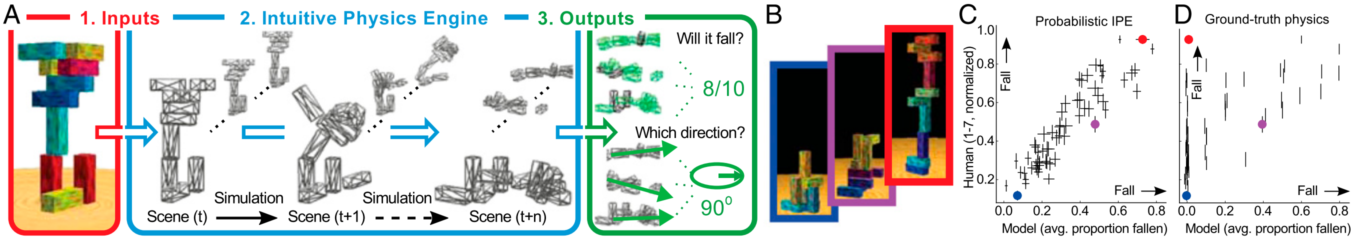

Fig. 2. (*A*) The IPE model takes inputs (e.g., perception, language, memory, imagery, etc.) that instantiate a distribution over scenes (*1*), then simulates the effects of physics on the distribution (*2*), and then aggregates the results for output to other sensorimotor and cognitive faculties (*3*). (*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.

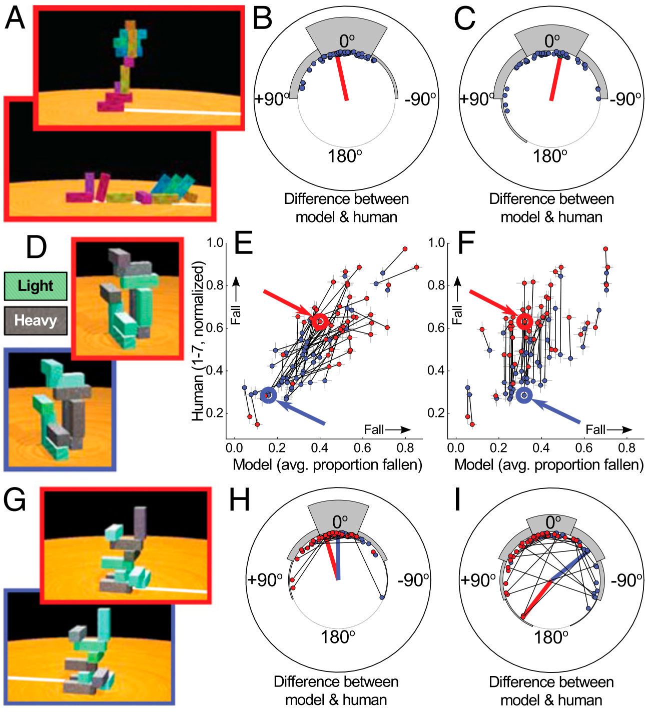

Fig. 3. (*A*) Exp. 2 (In which direction?). Subjects viewed the tower (*Upper*), predicted the direction in which it would fall by adjusting the white line with the mouse, and received feedback (*Lower*). (*B*) Exp. 2: Angular differences between the probabilistic IPE model's and subjects' circular mean 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 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

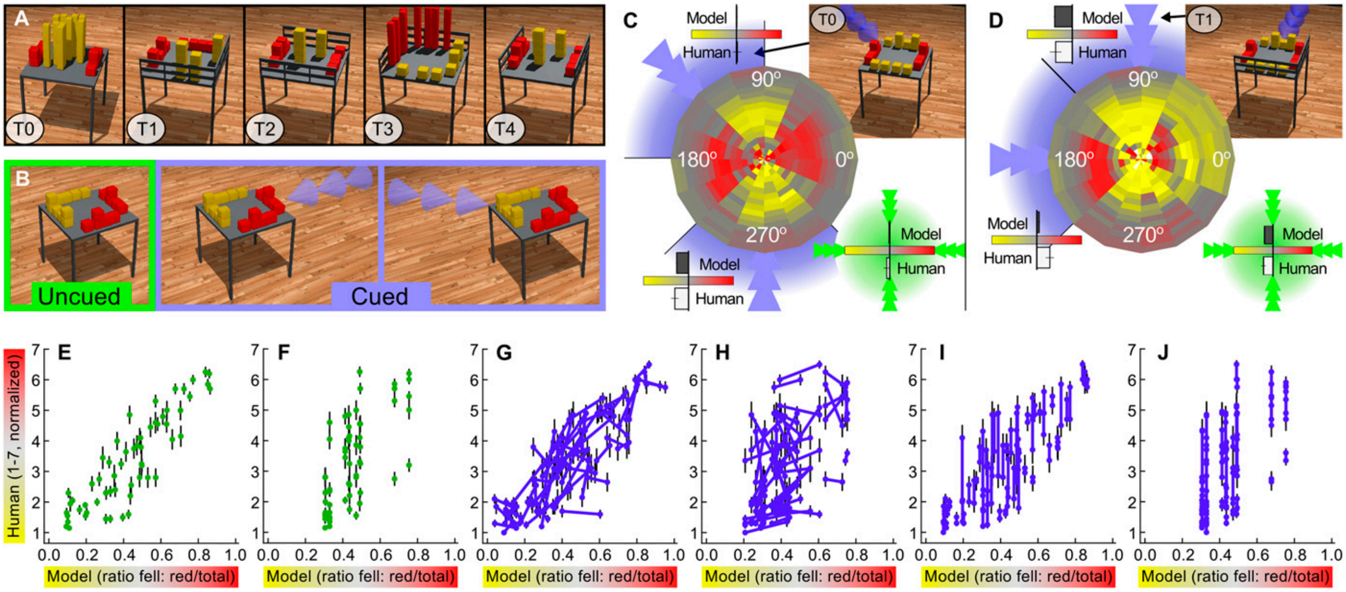

Fig. 4. Exp. 5 (Bump?). (*A*) Scene stimuli, whose tables have different obstacles (T0–T4). (*B*) In the uncued bump condition, subjects were not informed about the direction from which the bump would strike the scene; in the cued bump conditions, a blue arrowhead indicated the bump's direction. (*C*) 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.

Supporting Information

Battaglia et al. 10.1073/pnas.1306572110

SI Materials and Methods

Subjects and Stimulus Apparatus. All subjects volunteered in response to an advertisement posted on the Massachusetts Institute of Technology (MIT) Brain and Cognitive Sciences human subjects notification system (we did not collect background personal data, but our subject population is composed of roughly half MIT students or employees and half local community members). All gave informed consent, were treated according to protocol approved by MIT's Institutional Review Board, and were compensated $10/h for participation. All experimental sessions were 1 h long and took place in 1 d, and subjects ran in exactly one session in one experiment. All had normal or corrected-to-normal vision. Stimuli were presented on a liquid-crystal display computer monitor, which subjects free-viewed from a distance of 0.5–0.75 m. They indicated their responses by depressing a key on the keyboard or by adjusting the computer mouse and clicking to lock in their choice. The numbers of subjects per experiment are listed in Table S1.

Exp. 1. Subjects were presented with virtual, 3D tower scenes [rendered using Panda3D (1)], like those in Fig. 2*B* (main text) and Fig. S1, which contained 10 rectangular blocks (each with 3D aspect ratio 3:1:1). The blocks were stacked within a square column by a sequential random process such that when placed on the tower, no single block would fall off of its support (although, when later blocks were added, they might cause previously placed blocks to fall). This design increased the complexity of the scenes and judgments, by ensuring that pairwise comparisons of adjacent blocks would not provide information about whether the towers were stable or not. Subjects were asked, "Will this tower fall?", and responded on a 1–7 scale, in which 1, 4, and 7 corresponded to "definitely will not fall", "not sure", and "definitely will fall", respectively, with intermediate numbers indicating intermediate degrees of confidence. These stimuli were deliberately chosen to be challenging and to evoke judgments across this graded scale, to more clearly test the role of simulation and probability in physical scene understanding. For more natural everyday scenes, people's judgments and the intuitive physics engine (IPE) model's predictions would likely be more deterministic (more clustered at the endpoints of this 1–7 scale) and also more objectively accurate (more similar to ground truth physics).

On each trial, subjects first viewed the scene for a 3-s "prephysics" interval, in which gravity and other forces were absent (i.e., the scene remained static regardless of whether the blocks would fall when gravity was applied). During this interval the camera panned 180° around the tower so subjects could view the tower's full 3D geometry. (Pilot testing suggested that performance would be similar, if more variable, for subjects presented with a single static view.) At the end of this interval a cylindrical occluder dropped vertically over the tower and the subject was able to make a judgment under no time constraint. In Exp. 1 (feedback), feedback was presented immediately after the subject responded in the form of a 2-s movie in which gravity was turned on and either the tower remained upright or some (or all) of its blocks fell; the floor also changed color to distinguish between "fell" (red) and "remained standing" (green). In Exp. 1 (no feedback), the next trial began immediately after the subject responded.

Before the test session in Exps. 1–4, participants performed a 20-trial familiarization session with a set of tower stimuli different from those in the test session and with feedback, to acclimate them to the task timing and judgment, the response keys, and the parameters of the scene (e.g., the towers' general appearances and physical characteristics such as friction, etc.). Exp. 1's test session contained 360 trials: 60 towers (half of which would fall under gravity), repeated six times each across different trial blocks with randomized stimulus orderings and randomized trial-by-trial tower block colors and initial camera angles.

Exp. 2. Exp. 2 was similar to Exp. 1, except that all towers always fell under gravity (at least two blocks dropped). Subjects were asked "Which direction will the tower fall in?" and reported their judgments using the computer mouse by adjusting the orientation of a line on the floor, which extended from the base of the tower to the floor's perimeter, to indicate their expectation of its dominant fall direction (Fig. 3*A*, main text). Exp. 2's test session contained 360 trials, consisting of 60 stimuli repeated six times each in separate trial blocks, with the same randomizations of stimulus ordering, block coloring, and initial camera angles as in Exp. 1.

Exps. 3 and 4. Exps. 3 and 4 used a new set of tower scenes (Fig. 3 *D* and *G*) that was similar to that in Exps. 1 and 2, except that half of the blocks were 10 times heavier than the others. Blocks of different masses were visually distinguished by a dark, stone-like texture (heavy) and a pale green striped texture (light). As in Exps. 1 and 2, subjects were asked Will the tower fall? (Exp. 3) and Which direction will the tower fall in? (Exp. 4). The towers were organized into "state pairs": two towers whose geometries (block positions and poses) were identical, but whose heavy/light assignments were different. For some state pairs, the different mass assignments caused very different outcomes with respect to the model's predictions. As in Exp. 1, half of the towers in Exp. 3 would fall under gravity; as in Exp. 2, Exp. 4's towers always fell under gravity. In both experiments there were 48 towers with unique arrangements of blocks, each with two state-pair assignments of heavy and light blocks, for a total of 96 stimuli. The test sessions each contained 384 trials, consisting of the 96 stimuli repeated 4 times each in separate trial blocks, with the same randomizations as in Exps. 1 and 2.

Exp. 5. Exp. 5's scenes depicted a table on which a collection of blocks were arranged (Fig. 4*A*), half of which were red and the other half of which were yellow. Subjects were asked, "If the table were bumped, which color would be more likely to fall off?", and responded on a 1–7 scale in which 1, 4, and 7 corresponded to "definitely yellow", "not sure", and "definitely red". Subjects were instructed to assume that the bump's force was great enough to knock off at least one block. There were 12 possible block configurations that varied in the numbers and shapes of the blocks and the numbers, heights, and positions of the stacks in which they were arranged. There were also five different tables, one with a flat surface and four others with two short walls rigidly attached to different edges, which were designed to increase the complexity of the ensuing dynamics once the table was bumped. The 60 scene stimuli included all possible combinations of the 12 block configurations and five tables. We ran two conditions that differed by what information was provided about the ensuing bump: In the "uncued" condition, no additional information was provided about the bump; in the "cued" condition, a blue arrowhead in the scene pointed in the direction from which the bump would strike (Fig. 4*B*). Each stimulus was shown six times (with randomized red/yellow group assignments and viewing angles), two times in the uncued condition and four times in the cued condition (two times each with two different bump directions), for a total of 360 trials. Each block of 60 trials showed the stimulus from only one of the two conditions, ordered “no cue”, “cue”, “cue”, “no cue”, “cue”, “cue”. On each trial, the stimulus interval was 3 s, during which the camera panned over a 90° arc; the scene remained unoccluded during the response interval. After every fifth trial, feedback was presented in which a force was applied to the table for 200 ms with random magnitude but strong enough to cause at least one block to topple and fall off the edge. In the bump cue condition, the feedback’s actual bump matched the cued direction; in the uncued direction the bump’s direction was randomized. Subjects first performed a 20-trial familiarization session with feedback on each trial to become acquainted with the task and the effects of the bump. The familiarization stimuli contained different scenes with fewer blocks in simpler arrangements, so that the judgments were not challenging.

Exp. S1. Exp. S1 was almost identical to Exp. 1, except that the towers were selected such that all were of the same height. There were 432 experimental trials: 108 towers, repeated four times each, and trials were divided into four trial blocks, with the same randomizations over trial orderings, block colorings, and initial camera angles as in Exp. 1. Correlations between subjects’ judgments and IPE model predictions are shown in Fig. S4.

Exp. S2. Exp. S2 was identical to Exp. 2, except that subjects were asked to judge “How far will the furthest block come to rest?”. They reported their judgments using the computer mouse by adjusting the radius of a circle on the floor, centered at the tower’s base, to match the distance they expected the tower’s blocks to come to rest. The model’s output predicate was defined as the farthest point from the tower’s base on any block.

Data Analysis. In Exps. 1, 3, and 5, where people responded on a 1–7 scale, these judgments were rescaled linearly to lie between 0 and 1 and averaged across repetitions and subjects. We compared the model’s responses with subjects’, using a Pearson correlation. In Exps. 2 and 4 we computed the circular mean of subjects’ directional responses across repetitions and subjects. For Exps. 2 and 4, because judgments were angular, we used circular correlations (2) to quantify their agreement; note that circular correlation coefficient values are analogous but not directly comparable to Pearson correlations.

All correlation coefficients and 95% confidence intervals (CIs) were estimated using a bootstrap resampling procedure with 10,000 resamplings. The coefficients were taken as the median of the bootstrapped distribution; the lower and upper confidence intervals were always the 2.5th and 97.5th percentiles. All *P* values were estimated using a direct bootstrap hypothesis test over the 10,000 resamples (3).

In Exps. 2 and 4, for some towers, the model’s distribution of predicted directions was dispersed broadly around the circle and thus the model’s circular mean estimate was very unstable (single samples could cause the response to shift by up to 180°) (Fig. S2). In these cases the model was indecisive and its point estimates of the direction of fall were not meaningful. We thus excluded those towers for which less than 80% of the model’s distribution of predicted directions fell in any one-half of the circle, which left 47 of 60 towers included in the analysis for Exp. 2 and 72 of 96 for Exp. 4 (Fig. S2). In Exp. S2 we computed the mean of subjects’ radial responses across repetitions and subjects.

The reported *F*-scores comparing IPE model confidences in Exp. S2 and Exp. 1 refer to a one-way ANOVA *F*-test statistic, measuring how separable the model’s distributions of predictions were across stimuli within each experiment. They were computed as the ratio of the variance of per-stimulus prediction means over the means of the per-stimulus prediction variances.

SI Appendix: IPE Model

Computational Theory. The model forms a judgment, J_q, by computing the expected value of a physical property query, $\mathcal{Q}_q(S_{0:T})$, which is a function of the initial state (S_0) and the sequence of future states $(S_{1:T})$. The model’s knowledge about these states is summarized as the Bayesian posterior probability distribution given observed information about the states and latent forces $(I_{S_0:S_T}$ and $I_f)$.

Definitions.

- S_t : Scene state at time, t.
- $S_{t_0:t_1} = (S_{t_0}, S_{t_0+1}, \ldots, S_{t_1-1}, S_{t_1})$: Sequence of scene states from time t_0 to t_1.
- f_t : Extrinsic force applied beginning at time t.
- $f_{t_0:t_1}$: Sequence of extrinsic forces applied from time t_0 to t_1.
- I_{S_t} : Observed information about S_t.
- I_f : Observed information about $f_{t_0:t_1}$.
- $\psi(\cdot)$: Deterministic physical dynamics from t_0 to t_1, which maps S_{t_0} to a new state at time $S_{t_1} : S_{t_1} = \psi(S_{t_0}, f_{t_0}, t_1 - t_0)$. The force, f_{t_0}, is applied for a duration $t_1 - t_0$. The dynamics can be applied recursively,

$$S_{t_2} = \psi\left(\psi\left(S_{t_0}, f_{t_0}, t_1 - t_0\right), f_{t_1}, t_2 - t_1\right).$$

- We denote the repeated application of $\psi(\cdot)$ from $(t_0 : t_n)$ as $\Psi(\cdot)$,

$$\begin{aligned} S_{t_n} &= \psi\left(\ldots\psi\left(S_{t_0}, f_{t_0}, t_1 - t_0\right), \ldots, f_{t_{n-1}}, t_n - t_{n-1}\right) \\ &= \Psi\left(S_{t_0}, f_{t_0:t_{n-1}}, t_0 : t_n\right). \end{aligned}$$

- $\mathcal{Q}_q(\cdot)$: Output predicate corresponding to a query, q, which maps an initial (S_0) and a future sequence of scene states $(S_{1:T})$ to a judgment, $J_q = \mathcal{Q}_q(S_{0:T})$. In our experiments, the queries were sensitive only to the initial and final scene states (i.e., for Will the tower fall?, the query reflected how many blocks dropped from $t=0$ to $t=T$), and so $J_q = \mathcal{Q}_q(S_0, S_T)$.

Inputs. The model represents knowledge of S_t, using a probability distribution, $\Pr(S_t)$, which, if I_{S_t} is available, will be a posterior distribution defined by Bayes’ rule, $\Pr(S_t|I_{S_t}) = \frac{\Pr(I_{S_t}|S_t)\Pr(S_t)}{\Pr(I_{S_t})}$. Similarly, the model represents knowledge of $f_{t_0:t_1}$, using a probability distribution, $\Pr(f_{t_0:t_1})$, or, if I_f is available, $\Pr(f_{t_0:t_1}|I_f)$. For brevity, the remaining formulas assume observed information is available.

Physical inference. The probability of a future state, S_{t+1}, given a previous state, S_t, $\Pr(S_{t+1}|S_t, f_t) = 1$, for $S_{t+1} = \psi(S_t, f_t, 1)$, and 0 for any other value of S_{t+1}, because $\psi(\cdot)$ is deterministic. The model represents the distribution over initial and future states, S_0 and $S_{1:T}$, as determined by physics:

$$\begin{aligned} \Pr\left(S_{0:T}|I_{S_0}, I_f\right) = &\int_{f_{0:T-1}} \Pr(S_T|S_{T-1}, f_{T-1}) \cdots \Pr(S_1|S_0, f_0) \\ &\Pr(S_0|I_{S_0})\Pr\left(f_{0:T-1}|I_f\right) \mathrm{d}f_{0:T-1} \\ = &\int_{f_{0:T-1}} \Pr(\psi(S_{T-1}, f_{T-1}, 1)|S_{T-1}, f_{T-1}) \cdots \\ &\Pr(\psi(S_0, f_0, 1)|S_0, f_0)\Pr(S_0|I_{S_0}) \cdots \\ &\Pr\left(f_{0:T-1}|I_f\right) \mathrm{d}f_{0:T-1}. \end{aligned} \tag{S1}$$

In this work we focus on the influence that the latent forces, $f_{0:T-1}$, have on future scene states—not on the latent forces themselves—which is why they can be integrated out.

When only S_0 and S_T are required, as is the case in the present work, the intermediate times can be integrated out,

$$\begin{aligned}\Pr\left(S_T,S_0|I_{S_0},I_f\right) &= \int\limits_{f_{0:T-1}}\int\limits_{S_{1:T-1}} \Pr(S_T|S_{T-1},f_{T-1})\ldots\Pr(S_1|S_0,f_0)\\ &\quad \Pr(S_0|I_{S_0})\Pr\left(f_{0:T-1}|I_f\right)\mathrm{d}S_{1:T-1}\mathrm{d}f_{0:T-1}\\ &= \int\limits_{f_{0:T-1}} \Pr(S_T|S_0,f_{0:T-1})\Pr(S_0|I_{S_0})\Pr\left(f_{0:T-1}|I_f\right)\mathrm{d}f_{0:T-1}\\ &= \int\limits_{f_{0:T-1}} \Pr(\Psi(S_{t_0},f_{0:T-1},0:T)|S_0,f_{0:T-1})\\ &\quad \Pr(S_0|I_{S_0})\Pr\left(f_{0:T-1}|I_f\right)\mathrm{d}f_{0:T-1}.\end{aligned} \qquad \textbf{[S2]}$$

Note that because $\psi(\cdot)$ can be applied recursively,

$$\Pr(S_T|S_0,f_{0:T-1}) = \Pr(\Psi(S_0,f_{0:T-1},0:T)|S_0,f_{0:T-1}).$$

Outputs. The model's output judgment for a query, q, is

$$\begin{aligned}J_q &= \mathrm{E}\left[\mathcal{Q}_q(S_{0:T})|I_{S_0},I_f\right]\\ &= \int\limits_{S_{0:T}} \mathcal{Q}_q(S_{0:T})\Pr\left(S_{0:T}|I_{S_0},I_f\right)\mathrm{d}S_{0:T},\end{aligned} \qquad \textbf{[S3]}$$

where $\Pr(S_{0:T}|I_{S_0},I_f)$ is defined in Eq. **S1**.

Again, as our experimental queries were sensitive only to S_0 and S_T,

$$\begin{aligned}J_q &= \mathrm{E}\left[\mathcal{Q}_q(S_0,S_T)|I_{S_0},I_f\right]\\ &= \int\limits_{S_T}\int\limits_{S_0} \mathcal{Q}_q(S_0,S_T)\Pr\left(S_T,S_0|I_{S_0},I_f\right)\mathrm{d}S_0\ \mathrm{d}S_T,\end{aligned} \qquad \textbf{[S4]}$$

where $\Pr(S_T,S_0|I_{S_0},I_f)$ is defined in Eq. **S2**.

Experimental IPE Implementation. ***Inputs.*** As described in the main text, the IPE model tested experimentally had three numerical parameters, (σ,μ,ϕ), which controlled the state uncertainty, physical attributes, and latent force inputs to the simulation, respectively. They were defined and implemented as follows.

The model's state representation could be separated into the geometric state, G, and the physical state, p: $S=(G,P)$. In our experiments, G included the numbers, positions, poses, and shapes of the objects' states; and P included the mass density of each object, represented as the vector m, in addition to other physical parameters such as coefficients of friction and elasticity that were fixed across all studies to values typical for everyday wooden blocks. Similarly, $I_S=(I_G,I_P)$, where I_G represents observed information about G, e.g., the objects' visually indicated geometry in the image; and I_P represents observed information about P, e.g., the objects' visually indicated mass assignments in Exps. 3 and 4.

Because subjects viewed scenes from many viewpoints over a continuous 180° span, and these viewpoints were randomly chosen for each trial, we represented the model's visual inference of the initial scene geometry, G_0, as a simple, one-parameter viewpoint-invariant approximation to the Bayesian posterior,

$$\Pr(G_0|I_{G_0}) \approx \pi\left(G_0;\overline{G}_0,\sigma\right),$$

which represents the distribution over G_0 given the true geometry, $\overline{G}_0$, and the parameter, σ, which represents the magnitude of the posterior state uncertainty. The $\pi(\cdot)$ was defined by taking the true geometry, $\overline{G}_0$, and adding horizontal, zero-mean Gaussian noise (SD σ) to the ground truth object positions independently. Because the noise could cause interobject penetrations, the objects' coordinates were then transformed by a deterministic constraint-satisfaction procedure that selected the nearest configuration for which no objects violated each others' volumes. This procedure ran very small time steps of the physics engine, resetting the objects' velocities to zero after each step, which caused all that were detected as being in collisions to shift apart until no collisions were detected. We evaluated the plausibility of this viewpoint-invariant approximation $(\pi(\cdot))$ by comparing its samples with those of a prototype Bayesian vision system that we developed, which used Markov chain Monte Carlo (MCMC) to sample directly from the Bayesian posterior of scene geometries conditioned on specific subsets of images, under a likelihood function defined by a graphics rendering package (*SI Appendix: Bayesian Vision System*). Results were highly similar, in terms of both the samples of scene geometries (Fig. S8 *A–F*) and the predictions on our experimental tasks (Fig. S8 *G* and *H*). In our simulations, σ could take 1 of 11 possible values,

$$\sigma \in \{0.0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.5\}$$

(where the short width of one of the tower's blocks in Exps. 1–4 was 1, and the cube-shaped block's width in Exp. 5 was 2).

The model's representation of objects' mass densities m was controlled by the parameter μ. In Exps. 3 and 4, which contained heavy and light blocks, μ represented the scalar ratio between their respective mass densities; in Exps. 1, 2, and 5, all blocks had the same density, so $\mu=1$. The information provided by I_P was approximated as deterministically indicating which blocks were heavier (in Exps. 3 and 4) or that all had the same density (in Exps. 1, 2, and 5), and the model then used its assumption about μ to set each block's individual mass density. In Exps. 1, 2, and 5, $\mu=1$; in Exps. 3 and 4 it could take 1 of 12 values,

$$\mu \in \{0.25, 1.0, 2.0, 2.5, 3.2, 4.0, 5.0, 6.3, 8.0, 10, 13, 16\}.$$

For tower scenes (Exps. 1–4), all objects in the scene were adjusted so that the total mass was 2 kg. For Exp. 5 each block was 1 kg.

The model's latent force dynamics, $f_{0:T-1}$, represented possible vibrations and bumps that could be applied to the scene (main text). We approximated them as a horizontal force [angle θ and magnitude ϕ (main text)], applied from $t=0$ ms to $t=200$ ms (and no force after 200 ms) to the surface on which the objects were situated; for simplicity we drop the $0:T-1$ subscript and refer to $f_{0:T-1}$ as, $f=(\theta,\phi)$. The observation about the forces could be separated into two terms, $I_f=(I_\phi,I_\theta)$, where I_ϕ reflected language-based instructions like "If the table were bumped…" (Exp. 5). When available (Exp. 5, bump cue condition), I_θ represented the cue's indicated latent force direction; the model assumed θ was uniform over the range $[I_\theta-45, I_\theta+45]$. When I_θ was unavailable, the model assumed θ was uniform over the range [0, 360].

In Exps. 1–4, ϕ could take 12 possible values,

$$\phi \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.5, 2.0, 2.5, 3.0\},$$

and in our analyses we fitted the best value from Exp. 1. Fig. S3 shows the correlations between model and human judgments for a range of σ and ϕ values in Exp. 1. In Exp. 5 (the "Bump?" task), ϕ could take 16 values,

$$\begin{aligned}\phi \in \{&22.0, 25.6, 29.2, 32.8, 36.4, 40.0, 43.6, 47.2, 50.8, 54.4, 58.0,\\ &61.6, 65.2, 68.8, 72.4, 76.0\},\end{aligned}$$

where, in Exp. 5's scenes, $\phi=22.0$ usually caused only one or a few blocks to fall off the table, and $\phi=76.0$ caused many or almost all to fall off. The model's judgment took the expectation over all ϕ values.

Physical simulation. For our experiments, the model's judgment in Eq. **S4** was computed via a Monte Carlo approximation

$$J_q^{MC} \approx \frac{1}{k}\sum_{i=1}^{k} \mathcal{Q}_q\left((G_0,\mu)^{(i)}, \Psi\left((G_0,\mu)^{(i)}, f^{(i)}, 0:T\right)\right) \qquad \textbf{[S5]}$$

that used the Open Dynamics Engine (4, 5) to approximate $\Psi(\cdot)$, over a T that was always 2,000 ms with a step size of 1 ms, and where k is the number of independent simulation samples, initialized by independently drawing values of S_0 and f from the approximate posterior distributions as described above. In Exps. 1–4, $k=48$. In Exp. 5, $k=12$, and only the values of S_0 were drawn independently; rather than sampling f, we computed predictions for 16 equally spaced values of θ, crossed with the 16 values of ϕ (Fig. 4 *C–E*).

Query definitions. All queries regarded an object as having "fallen" if, between S_0 and S_T, its z coordinate was displaced downward greater than 0.025 unit. For Exp. 5, objects were defined as having "fallen off" the table if their z coordinate was less than 0 at time T (the table's surface was at $z=0$).

The queries corresponding to the four experimental tasks were as follows:

- Exps. 1 and 3 (Will it fall?): $\mathcal{Q}_{fall}(S_0, S_T)$ = the fraction of objects that fell.
- Exps. 2 and 4 (In which direction?): $\mathcal{Q}_{dir}(S_0, S_T)$ = the angle of the center of mass of the fallen blocks, in the x,y plane. For $\mathcal{Q}_{dir}$'s Monte Carlo sum, we used the circular mean across the angular query outputs.
- Exp. 5 (Bump?): $\mathcal{Q}_{bump}(S_0, S_T)$ = the ratio of red to total blocks that fell off the table.
- Exp. S2 (How far?): $\mathcal{Q}_{far}(S_0, S_T)$ = the radius of the farthest fallen block from the base of the tower, in the x,y plane.

Due to ambiguities and vagueness inherent in language, more than one output predicate could be consistent with any given task query. For example, although we represented answers to the Will it fall? query as the average proportion of blocks that fell, other possibilities (e.g., the proportion of simulations for which one or more blocks fell or the proportion for which more than half of the blocks fell) would give similar results.

SI Appendix: Model-Free Accounts

An alternative explanation for people's physical scene understanding is that they do not possess some model of the world, but instead use model-free methods that depend heavily on their experienced interactions with the world. People have been exposed to many stable and unstable arrangements of objects over their lifetime, and perhaps when presented with a new scene such as our experimental stimulus, they consult some stored representation of similar scenarios they have previously experienced and produce a response that reflects the outcomes of those scenarios. Such stored representations might be exemplar based, composed of individual examples of scene and outcome pairs, or feature based, encoding their experiences as compressed features—the gross shape of the scene, the number of vertical elements, etc.—and a function that takes as inputs these features and returns as output statistically dependent physical outcomes. Either the features or the parameters of this input–output function (or both) could be constructed on the basis of experience. To what extent could one of these model-free methods explain subjects' performance in our experiments?

One reason to be skeptical of a model-free account is the lack of significant practice effects over the course of our experiments (*SI Appendix: Analysis of Learning*), which might be expected if people's judgments were driven primarily by data-driven learning mechanisms.

However, the central argument for a model-based probabilistic physical simulation mechanism over a primarily model-free mechanism based on learning from experience is the former's generality. Although the same IPE model could be applied to each new task by using a predicate that was appropriate for the query and to each new scene type by adjusting its parameters to reflect the objects' physical attributes and the scene's patterns of applied forces, the extensive amount of variation across natural settings means that no compact set of trained features or exemplars is even applicable to all tasks or scenes.

Still, to formally evaluate an account based on memory or learned features, we built separate feature models for the Will it fall?, How far?, In which direction?, and Bump? tasks and fitted their parameters to best predict subjects' data in each experiment, using ridge regression (a linear model that prevents overfitting by subjecting the coefficients to an L2 penalty). The features used are listed in Table S2. For each experiment, the penalty parameter was selected through a cross-validated fitting procedure on subjects' data. This method is relatively generous to a feature-based account because it allows the feature weights to vary arbitrarily across experiments, taking on whatever values best fits people's performance in each particular experiment.

Because standard multivariate regression analysis methods are not available for circular data, we report the best individual circular correlation between people's responses and any single feature in the In which direction? tasks. In Exps. 3 and 4, some geometric features implicitly took the blocks' masses into account [e.g., $F_F(2)$, height of the center of mass]. To grant these features the ability to make judgments that were sensitive to the heavy/light assignments, we generated features for a range of assumed μ values and used those that best fitted people's judgments.

Across experiments, the IPE model fits were generally significantly better than those of the best feature-based models, often dramatically so (Fig. S4). This was true even allowing for features that were selected specifically for each task and multiple free parameters that were tuned to maximize their fits to each experiment separately.

The results of fitting the best feature-based account to each individual experiment were as follows. In Exp. 1, across the feedback and no-feedback conditions, one heuristic predictor, the tower's height [Table S2, $H_F(1)$], was best correlated with subjects' responses ($\rho=0.75$, 95% CIs [0.68, 0.81]), so we conducted a controlled variant (Exp. S1) identical to Exp. 1 (feedback) except with different subjects (n = 10) and 108 new towers (each repeated over four blocks) that were all of the same height, to assess performance when the most dominant geometric heuristic was neutralized. The IPE, with parameters identical to those of Exp. 1 ($\sigma=0.2$ and $\phi=0.2$), again had a significantly higher correlation with people ($\rho=0.81[0.74, 0.87]$) than the geometric heuristics ($\rho=0.70[0.59, 0.78]$, $P<0.001$). In Exp. 2, the best geometric feature had a significantly lower circular correlation with people's circular-mean responses than the IPE ($\rho_{circ}=0.39[0.21, 0.56]$, $P<0.001$). In Exp. 3, the regression-fit mass-sensitive feature predictions (with best-fit $\mu=6$) had a correlation of $\rho=0.71[0.61, 0.79]$ and the mass-insensitive features (with $\mu=1$) had a correlation of $\rho=0.61[0.46, 0.72]$, which were also both significantly lower than that of the IPE model ($P<0.02$ and $P<0.001$, respectively). The features' state-pair differences (with $\mu=6$) correlation, $\rho=0.54[0.34, 0.73]$, was also significantly lower ($P<0.03$) (the mass-insensitive features did not make different state-pair predictions). In Exp. 4, the best mass-sensitive feature [$F_D(6)$, best-fit $\mu=16$] and mass-insensitive feature [$F_D(6)$, $\mu=16$] had correlations of $\rho_{circ}=0.43[0.28, 0.58]$ and $\rho_{circ}=0.31[0.17, 0.45]$, respectively, which were both significantly lower than that of the IPE model ($P<0.001$ and $P<0.1$). However, the state pair circular correlation between the IPE model and the subjects was not significantly higher ($P<0.07$) than the feature correlation

($\rho_{circ} = 0.50[0.18, 0.75]$ vs. $\rho_{circ} = 0.25[0.066, 0.45]$, respectively). In Exp. 5, the regression-fit features had significantly lower correlations than the IPE model did with people's responses, in both the uncued and the cued conditions, respectively ($\rho = 0.73[0.58, 0.85]$, $P < 0.01$; $\rho = 0.68[0.58, 0.77]$, $P < 0.001$).

We do not claim that geometric features learned through experience play no role in physical scene understanding: In one experiment, Exp. S2 (How far?), a simple feature was significantly better than the IPE model at predicting people's judgments. Below (*SI Appendix: Approximations*) we further discuss why and when features may be used instead of mental simulations. However, we argue against the sufficiency of a purely memory- or feature-based account and in favor of a general capacity for simulation, based on the IPE model's distinctive ability to explain quantitatively the whole set of experimental results presented here, as well as people's inferences across a wide range of real-world scenes and tasks such as those in Fig. 1 and the examples discussed in main text under *Architecture of the IPE*. It is difficult to imagine a set of features flexible enough to capture all of these inferences, yet compact enough to be learnable from people's finite experience.

SI Appendix: Approximations

Our IPE model as tested so far is surely incomplete in key ways. It is likely that people's capacity to represent the full physical state of a scene, to simulate the dynamics of many objects in motion over time, and to maintain faithful representations of probabilities is more limited–and also more adaptive—than our use here of a simple physics engine running several dozen stochastic samples. The human IPE likely trades off precision for efficiency much more aggressively than engineers typically do. People may summarize the individual objects in a many-object scene coarsely, substituting in aggregrate representations that approximately capture how such "stuff" tends to behave. They may use simplified shape representations, such as spheroids or convex polyhedra, that allow multiobject interactions to be computed more efficiently. The temporal resolution of the IPE's simulations may be low, potentially leading to errors and biases that only more finely spaced time steps would be able to avoid. The laws of physics embedded within the human IPE may not uphold basic physical principles, such as the conservation of energy and momentum, and the simulation may blend mechanics principles that are objectively independent into combined rules; e.g., because friction is ever present, objects in simulated motion may gradually but constantly lose energy and momentum over time. The IPE may represent uncertain sets of possible states with only a few samples or instead represent a state's probability, using a continuous-valued weight rather than through its frequency of occurrence.

These approximations may have little effect in many everyday contexts, where even short simulations based on coarse representations of objects' shapes and positions are sufficient to make useful predictions. However, they could result in perceptual errors or illusions with more complex scenes or more demanding tasks, pointing to ways in which our model could be improved to better capture the approximations and shortcuts people exploit. The following subsections illustrate in more depth several types of approximations that the human IPE may make and the role they might play in people's physical intuitions.

Approximating Physics. One kind of approximation is motivated by the intrinsic difficulty of making certain kinds of judgments via simulation in the presence of complex dynamics. The Open Dynamics Engine (ODE) and other standard physics engines can simulate highly nonlinear systems, such as many-body collisions, for which accurate predictions over even short time intervals are computationally intensive and probably beyond what the human IPE can perform. Consider a bowling ball at the moment it leaves the bowler's hand: We can mentally extrapolate its path for several seconds as it rolls down the lane, but the instant it strikes the pins, the ensuing motions become unimaginably complex. The same dynamics apply when a stack of blocks falls or is knocked over, as in Exps. 1–5, but did not pose serious challenges there because the judgments queried were mostly insensitive to collisions occurring beyond the early stages of the simulation. To test whether people might rely on alternatives to simulation in cases where these factors matter more, we conducted a supplementary study (Exp. S2), using the same tower stimuli of Exps. 1–2, but instead asked subjects to predict "How far will the blocks come to rest?". This judgment depends sensitively on tracking each block precisely until it comes to rest, often after multiple collisions. Here the IPE model makes much less confident predictions (*F*-ratio of 66.34 vs. 172.17 in Exp. 1, where higher values indicate greater separation between the model's simulated distributions across stimuli) (*SI Appendix: Data Analysis*), and people's judgments on this task were also much less accurate, as assessed both by correlation with ground truth ($\rho = 0.38$ vs. $\rho = 0.64[0.46, 0.79]$ in Exp. 1) and by correlation with the IPE model ($\rho = 0.71[0.56, 0.81]$ vs. $\rho = 0.92[0.88, 0.94]$ in Exp. 1 using the same parameter values) in Exps. 1 and 2. Intriguingly, in this case a simple geometric feature—the height of the tower—was correlated with the IPE model's inferences and better predicts people's judgments ($\rho = 0.93[0.87, 0.96]$) than the model does (Fig. S4). People may fall back on such learned features as simple heuristics for predictions when complex dynamics make mental simulation impractical, just as a bowler learns that targeting the 1 and 3 pins from a slight angle predicts a strike even though she cannot imagine how the pins will move to produce that outcome.

Approximating the Scene. Another kind of challenge for mental simulation can arise under even simple dynamics but with complex object shapes. In our IPE model, as in most standard physics simulation engines, predicting an object's response to gravity or other forces depends on representing its center of mass and moments of inertia. These variables are easy to estimate for the blocks in Exps. 1–5, with their cuboidal shape and uniform density, but for objects whose mass is distributed in an unknown way over a complex 3D volume, they likely exceed people's ability to estimate with much precision. People may use strongly simplifying priors, such as taking an object's mass to be uniformly distributed over a coarse approximation to its shape (e.g., a cuboid, an ellipsoid, or a convex hull). This would be consistent with recent reports that people can accurately judge the stability of asymmetric objects when they are convex (6), while also explaining why people incorrectly predict the stability and dynamics of objects with more complex, nonconvex shapes (Fig. S5): They expect a wheel rim will roll downhill at the same rate as a filled disk (Fig. S5*A*) (7), when in fact it rolls more slowly, and they are surprised to see the dragonfly in Fig. S5*B* apparently defying gravity, when in fact it is balanced stably around its center of mass (Fig. S5*C*).

Approximating Probabilities. There are several important questions of approximation in our proposed IPE's sample-based representation of probabilities. How many simulation samples does the human IPE use? How faithfully do these samples represent probabilistic quantities of interest? We explored these questions by examining the variance of subjects' responses, which will decrease with the number of simulation samples that contribute to each judgment, k (Eq. **S5**), in the same way that the (squared) SEM in experimental statistics will decrease as the sample size is increased. For each stimulus in each experiment, we computed the across-subject response variances (or circular variances, in Exps. 2 and 4) and model variances as follows. We first normalized the subjects' and model's responses in Exps. 1, 3, and 5 to be in [0, 1] so they would have common coordinates (note that Exps. 2 and 4's model and subject responses did not need to be normalized because they were both already in radians). We then shifted their values so that they would have a mean of zero across repetitions for each stimulus to remove the component of variance due to mean differences. Fig. S6 *A–E* shows the relationships between these model and subject response variances; the colored lines correspond to the relationships between the variances for different values of k (Fig. S6 legend for details). The relationship between the model and subject variances favors an effective value of $k=1$ for Exps. 1–4 and $k=5$ or 6 for Exp. 5. However, these estimates are based on the assumption that the sampling variance was the only source of trial-by-trial variance and thus represents a lower bound: There are surely other sources of variance that are typically present in psychophysical data, such as general decision-making noise. This may be why in Exps. 1–4 the subjects' variances are elevated above the $k=1$ line. To separate the sampling variance from these other stimulus-independent noise sources and better estimate the true value of k, we performed the following linear regression analysis. If we assume that the sampling variation is statistically independent of the other sources of variance, then the variance in people's responses is the sum of these two variance components and the slope of the fit line will correspond to $\frac{1}{k}$, whereas the intercept will correspond to ω^2, where ω is the SD of the other noise sources. The estimated k and ω values for Exps. 1–5, respectively, were

$$\begin{aligned}
k_1 &= 4.8[2.2, 66], \omega_1 = 0.27[0.25, 0.29]\\
k_2 &= 2.8[1.9, 5.1], \omega_2 = 0.56[0.51, 0.60]\\
k_3 &= 4.5[2.9, 8.9], \omega_3 = 0.21[0.19, 0.22]\\
k_4 &= 3.1[2.1, 5.4], \omega_4 = 0.47[0.42, 0.51]\\
k_5 &= 7.4[4.8, 13], \omega_5 = 0.09[0.07, 0.10].
\end{aligned}$$

This suggests that subjects typically formed their judgments based on 3–7 simulation samples per trial. A possible reason for the small differences between Exps. 1–4's k values of 3–5 and Exp. 5's of 7 is that Exp. 5's model predictions included uncertainty in the bump's direction and magnitude, in addition to the state uncertainty (σ) also present in Exps. 1–4: It may be that in such cases of greater uncertainty the IPE increases its numbers of samples to more accurately approximate the physical outcome. The differences between the ω terms in Exps. 1 and 3 and in Exp. 5 may reflect different effects that the tasks' demands caused in subjects' decision processes (Exps. 2 and 4's responses were in radians and so not comparable to the other experiments' ω terms).

That these estimated values of k were around 3–7 is consistent with the "one and done" account of Vul et al. (8), which claims that many judgments and behaviors might be the product of taking a small number of probabilistic samples from an internal posterior distribution. To examine how the number of samples affects the IPE model's ability to make accurate probabilistic inferences, for each experiment we created several IPE model variants that ran only a small number of samples and compared their judgments to those of the original IPE model whose judgments were based on many samples. Fig. S6 *F–J* shows that across experiments, the IPE model needs only a small number of samples to make predictions that correlate well with the predictions of the original IPE model; the correlations associated with model variants that used the (rounded) numbers of samples estimated above were 0.93, 0.83, 0.89, 0.84, and 0.93 for Exps. 1–5, respectively. This demonstrates that for our tasks, small numbers of samples are generally sufficient to make similar predictions to those of a model that has a more complete representation of probabilities.

Besides the simple Monte Carlo methods that we used here, there are other more sophisticated ways to use sampling in simulation to represent physical uncertainty over time. These are especially useful in cases when new observations are being continually collected (unlike in most of our experimental tasks, where all relevant observations are made at the beginning of each trial). Sequential importance sampling and particle filtering can use time-dependent samples that have weights associated with them, where the weight is proportional to an estimate of the posterior probability of that sample's state given the observations up to that point in time. The unscented Kalman filter uses a procedure to select a representative nonrandom subsample of states to which to apply the deterministic physical dynamics, to capture uncertainty over time. Future work should pursue this question of how exactly samples are used to represent and update dynamically changing probabilities within a simulation-based framework.

SI Appendix: Analysis of Learning

Across all experiments, our analyses treated the subjects' data as stationary and constant by collapsing across multiple stimulus repetitions, yet subjects performed hundreds of trials with repeated stimulus presentations and often received feedback. One possibility is that subjects arrived at or substantially improved their behavior through learning over the course of the trials, rather than drawing primarily on a fixed internalized model of physics. We examined each subject's responses for evidence of practice or learning effects by computing their judgments' differences from the IPE model and its variants as a function of trials completed (Fig. S7). Fig. S7 *A* and *B* shows subjects' time-averaged differences as a function of trial number for Exp. 1's feedback and no-feedback conditions, respectively; Fig. S7 *C* and *D* shows the judgment differences averaged across subjects. There appear to be at most negligible changes over the course of the trials for most individual subjects, as well as for the average over subjects. An almost-perfect correlation ($\rho=0.95$, 95% CIs [0.95, 0.95]) between subjects' mean responses in Exp. 1's no-feedback and feedback conditions (Fig. S7*E*) also suggests that feedback did little to alter people's judgments.

We also fitted regression lines to every subject's judgment differences as a function of trials completed and found that the largest shift in any single subject's judgments away from the model was +20%, whereas the largest shift toward any model variant was −17% (mean −1.6%, SD 6.6%). Fig. S7 *F–L* shows histograms of these slopes per experiment; most shifted by less than 10%, which indicates that subjects' responses were largely constant across trials. Fig. S7*M* shows these histograms (for shifts with respect to the probabilistic IPE model) pooled across all experiments and subjects, as well as a bootstrap resampled set of all experiments' data with the trial orderings randomized to express the hypothesis that there was no true shift in subjects' responses across trials. A bootstrapped hypothesis test of the difference between the empirically observed distribution of shifts and this no-shift reference distribution was not significant ($P=0.38$). Taken together, these analyses suggest that feedback played a minimal role and that there was little if any effect of practice or learning across trials.

SI Appendix: Bayesian Vision System

Our IPE model uses a simplified input representation of the scene: a sample of its 3D geometric state that approximates a Bayesian posterior distribution on scenes' given images. To explain physical scene understanding more fully, however, we ultimately need to capture all of the factors that govern how people infer underlying scenes from observed images, including effects of viewpoint, occlusion, and so on. Although a full treatment of the visual inference problem is beyond our scope here and would stretch the bounds of most conventional machine vision systems, here we offer an initial attempt both to validate our model's assumptions about the input geometry distribution and as a proof of concept to motivate future research integrating Bayesian vision with probabilistic physical reasoning.

We implemented an image-based vision system for approximating the Bayesian posterior distribution over G given I_G by MCMC in an inverse-graphics model and compared its judgments (Fig. S8 D and F) to those of our IPE's viewpoint-invariant, π-based samples (Fig. S8 C and E) for the tower stimuli in Exps. 1 and 2 (Fig. S8 A and B). The system's geometric state estimate G was defined as the positions and poses of all 10 blocks in the tower, and its observed evidence I_G was three 256×256 8-bit (256 grayscale levels) images of the tower rendered under perspective projection (using OpenGL) from a series of three viewpoints rotated by 45° and then filtered using a normalized 2D Gaussian blur kernel with x,y scale parameter values of 11 pixels. The blur kernel's scale was selected so that the variance of the Bayesian system's samples would roughly match the viewpoint-invariant model's variances. The prior over G was assumed to be uniform over all possible positions and poses of the blocks. The likelihood (probability of the image data given the geometric scene state) was defined as the product of per-pixel normal distributions (whose SD was 128 grayscale levels) between the observed image and an OpenGL-rendered image of the latent scene.

We used a Metropolis–Hastings (MH) sampling algorithm (9) to draw 5,000 geometric state samples (Fig. S8 D and F), with samples grouped into 10-sample scans. Within a scan, each block's state was updated once, and the order of block updates was chosen uniformly with replacement, independently per scan. The MH algorithm's proposals consisted of updating each block's horizontal position, drawing a value from a horizontal bivariate normal distribution (with a diagonal covariance matrix whose SD terms were 0.2), and adding that vector to the block's current position, keeping its vertical position fixed. We initialized the sampler's latent state at the true state to minimize "burn-in" overhead and additionally discarded the first 1,000 samples of each run as burn-in as well. These simplifications make the system not generally applicable for performing vision in arbitrary scenes. However, they allowed us to conveniently draw approximate posterior samples representing how people could parse our experimental stimuli, under the assumption that we can accurately perceive the numbers of objects and their general positions.

Fig. S8 G and H shows that when the Bayesian system's samples are input to the IPE model, the resulting judgments are highly correlated (Exp. 1, $\rho=0.90[0.84,0.94]$; Exp. 2 [circular], $\rho=0.94[0.91,0.97]$) with judgments of the viewpoint-invariant (π-based) model used in the main text.

1. CMU Entertainment Technology Center (2010) Panda3D. Available at www.panda3d.org. Accessed October 7, 2013.
2. Fisher N, Lee A (1983) A correlation coefficient for circular data. *Biometrika* 70(2):327–332.
3. Efron B, Tibshirani R (1993) *An Introduction to the Bootstrap* (Chapman & Hall/CRC, New York and London), Vol 57.
4. Smith R (2010) Open Dynamics Engine. Available at www.ode.org. Accessed October 7, 2013.
5. Baraff D (2001) Physically based modeling: Rigid body simulation. SIGGRAPH Course Notes, *Association for Computing Machinery SIGGRAPH* 2:2–1.
6. Cholewiak SA, Fleming RW, Singh M (2013) Visual perception of the physical stability of asymmetric three-dimensional objects. *J Vis* 13(4):12.
7. Proffitt DR, Kaiser MK, Whelan SM (1990) Understanding wheel dynamics. *Cognit Psychol* 22(3):342–373.
8. 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.
9. MacKay D (2003) *Information Theory, Inference, and Learning Algorithms* (Cambridge Univ Press, Cambridge, UK).

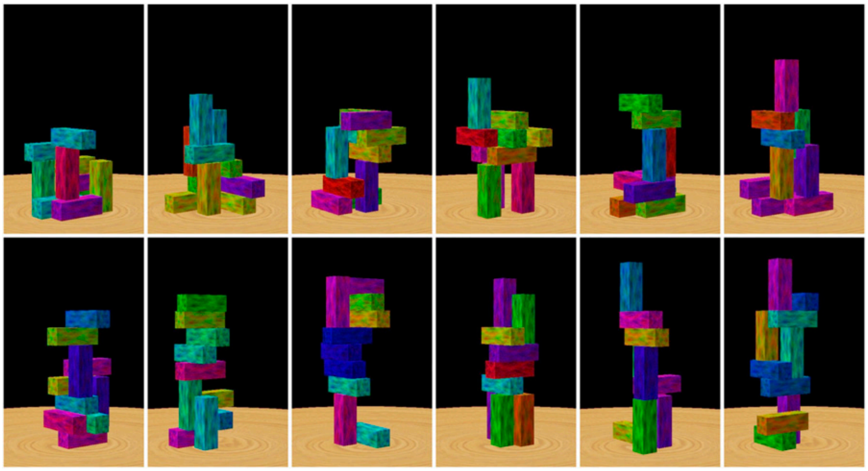

Fig. S1. Example of tower stimuli from Exp. 1.

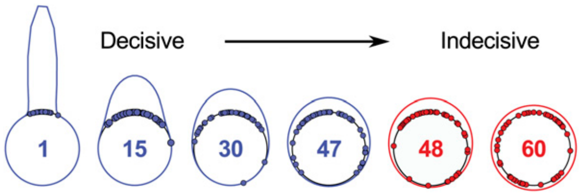

Fig. S2. Exp. 2 IPE model's decisiveness. Each ring plot shows a sample of one tower's model-predicted directions. The towers are sorted by how concentrated the predictions are (numbers indicate rank). The 47th tower and lower (blue) were included in the analysis but the 48th and higher (red) were not.

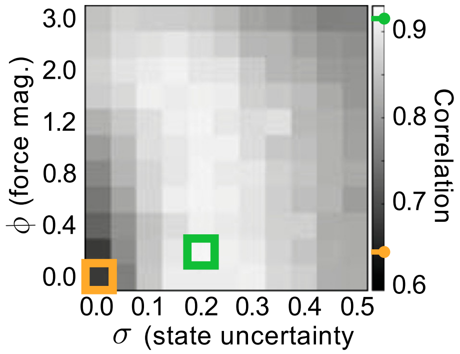

Fig. S3. Correlations between Exp. 1's model and human judgments as a function of IPE parameters, σ (*x* axis) and ϕ (*y* axis). The green and orange boxes/lines indicate the probabilistic and ground truth best fit parameters, corresponding to Fig. 2 *C* and *D*, respectively.

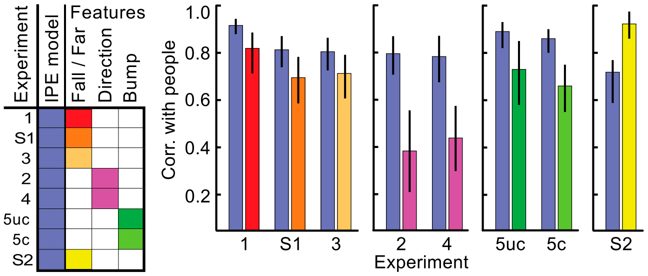

Fig. S4. Summary of model fits from Exps. 1–5 and Exps. S1 and S2. Bar graphs show the correlations (including 95% CIs) with people's judgments for both the IPE model and the best feature-based alternative account, across each experiment. Exps. 5uc and 5c refer to Exp. 5 uncued and cued conditions, respectively. Different features were required to fit behavior in different experiments. The graph (*Left*) shows which sets of features were applicable (colored boxes) or inapplicable (white boxes) to each experiment.

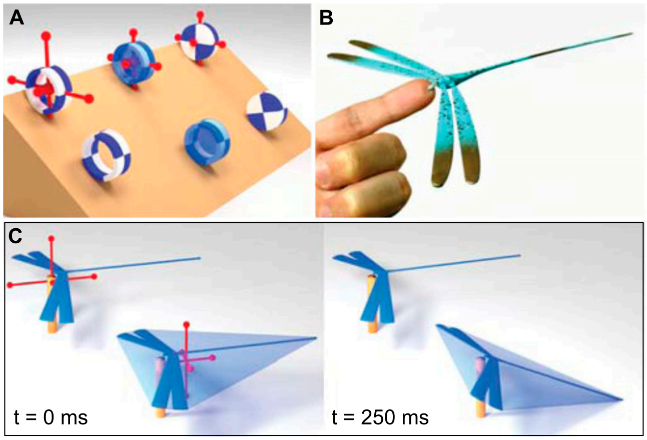

Fig. S5. Physics illusions and errors may arise from how the IPE approximates objects' dynamical properties. Moments of inertia along each object's principal axes are indicated by the length (inverse) of the red bars, centered at the estimated center of mass. (*A*) A wheel rim (leftmost object) has a different inertial tensor than a disk (rightmost object), such that it rolls downhill more slowly, but these inertial tensors are nontrivial to calculate. Treating each object's mass distribution as uniform over its convex hull gives a cognitively plausible approximation in general and here predicts (as people naively expect) that the wheel rim (middle object) has the same inertial properties and rolls at the same rate as the disk. (*B*) A toy dragonfly perches dramatically on its nose and is surprisingly balanced. (*C*) A model dragonfly with a physically correct center of mass and inertial tensor (upper) remains balanced after 250 ms in a simulation, but modeling it with the same convex-hull approximation as in *A* (lower) locates the dragonfly's center of mass slightly behind its nose (as people do) and leads to the intuitive expectation that it will tip from its perch.

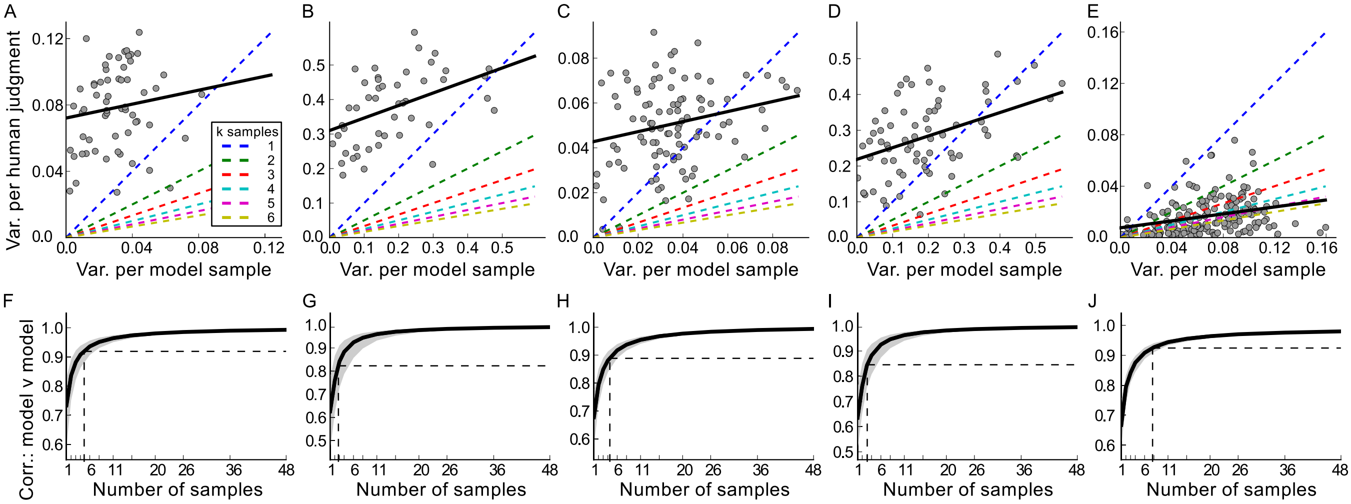

Fig. S6. Analysis of the number of IPE simulation samples. The columns of subplots correspond to Exps. 1–5, respectively. (*A*–*E*) Each point indicates the variance across human responses for a single stimulus (*y* axis) vs. the variance across model samples for that same stimulus (*x* axis). The blue dashed line (with the greatest slope) indicates a one-to-one correspondence between the model's sample variances and people's judgment variances, which would be consistent with each person's judgment being based on a single simulation sample. If instead each person formed judgments by taking a mean across $k > 1$ samples (drawn from the same probabilistic model), then we would expect the variances of people's judgments to be smaller than the model's sample variances by a factor of *k* (analogous to the squared SEM). The other dashed lines, with decreasing slopes ($\frac{1}{k}$, for $k = 2 \ldots 6$), depict expected correspondences between the human response variances and the model's sample variances if human judgments were based on means of 2–6 simulation samples, respectively. The black solid lines show best-fit regression lines, whose intercepts reflect stimulus-independent trial-by-trial variance due to sources other than sampling variability. (*F*–*J*) The thick black line depicts the correlation (*y* axis) between the IPE model's predictions based on *k* samples (*x* axis) and the model's predictions with the full set of samples from the original IPE; the gray ranges are 95% CIs (estimated by bootstrapped resampling of the *k* samples); at $k = 48$ samples the correlations converge to 1. The vertical dashed line indicates the best-fit value of *k* estimated from the linear regression analysis (*SI Appendix: Approximating Probabilities*), and the horizontal dashed line indicates the correlation level at this value of *k*.

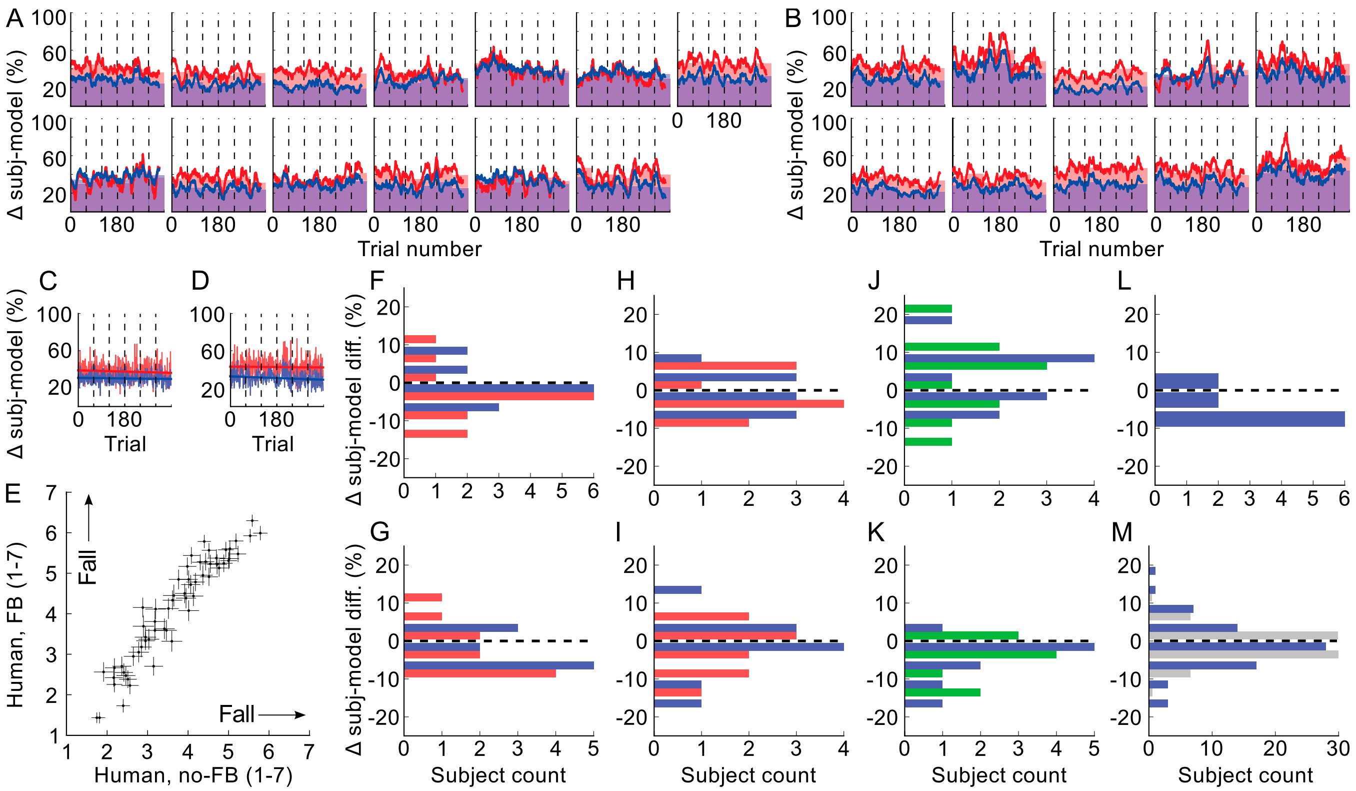

Fig. S7. Analysis of learning. (*A*) Individual subjects' average (over 20-trial sliding window) response differences (percentage, *y* axis) from ground truth (red: $\sigma = 0$, $\phi = 0$) and probabilistic IPE model (blue: $\sigma = 0.2$, $\phi = 0.2$) as a function of trial number (*x* axis) for Exp. 1's feedback condition. The vertical dashed lines indicate different trial blocks. (*B*) The same as *A* for Exp. 1's no-feedback condition. (*C*) Average response differences across all subjects (not a sliding window) for Exp. 1's feedback condition. The overlaid colored lines are best-fit regression fits. (*D*) The same as *C*, but for Exp. 1's no-feedback condition. (*E*) Correlation between subjects' average judgments (raw 1–7 responses, with SEM) in Exp. 1's no-feedback condition (*x* axis) vs. Exp. 1's feedback condition (*y* axis). (*F*) Histogram of slopes (*y* axis) of best-fit regression lines fitted to individual subjects' response differences from ground truth (red) and the probabilistic IPE model (blue) for Exp. 1's feedback condition. The slopes' units are change in response differences (percentage) across the experimental session; 0.0 indicates no change and −100% indicates a change from 100% error to 0% error. (*G*) The same as *F* for Exp. 1's no-feedback condition. (*H*) The same as *F* for Exp. 1 (same height). (*I*) The same as *F* for Exp. 2. (*J*) The same as *F* for Exp. 3. The blue and green bars are the slopes from fits to the mass-sensitive and mass-insensitive IPE models, respectively. (*K*) The same as *J* for Exp. 4. (*L*) The same as *F* for Exp. 5. The bars are slopes from fits to the full IPE model. *A*–*D* and *F*–*L* show minor changes in subjects' response differences from either the ground truth or the probabilistic IPE model across trials, indicating minimal effects of practice or learning. (*M*) Histograms of all slopes across experiments with respect to the probabilistic IPE model (blue) and the (rescaled) bootstrapped distribution of slopes for subjects' data with randomly permuted trial orderings (gray).

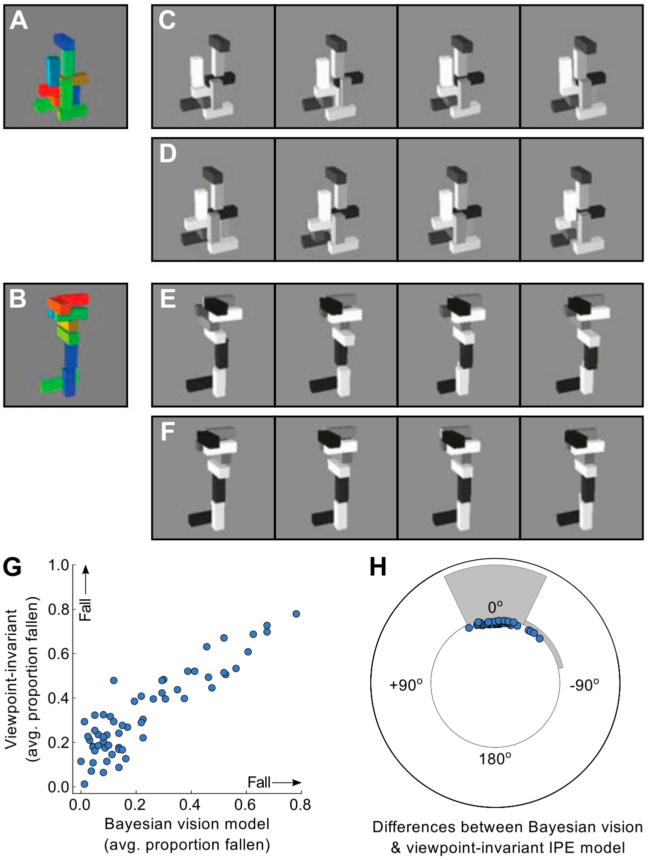

Fig. S8. Bayesian vision system. (*A* and *B*) Two original tower images. (*C* and *E*) Viewpoint-invariant (π-based) scene samples. (*D* and *F*) Bayesian vision system's scene samples. (*G*) Correlation between Exp. 1's model judgments for IPE model variants that input the Bayesian vision system's scene samples (*x* axis) and the viewpoint-invariant (π) samples used in the main text (*y* axis). (*H*) The same as *G*, but for Exp. 2.

Table S1. Numbers of subjects per experiment

Experiment no.	Judgment	No. subjects
1, feedback	Fall?	13
1, no feedback	Fall?	10
S1, same height	Fall?	10
2	Direction?	10
3	Fall?, mass	11
4	Direction?, mass	10
5	Bump?	10
S2	Far?	10
Total		74

Table S2. Geometric features

$F_F(\cdot)$	Exps. 1 and 3 and Exps. S1 and S2, Will it fall?/How far?
$F_F(1)$	Tower’s height
$F_F(2)$	Height of the tower’s center of mass
$F_F(3)$	Minimum critical angle[*] of the tower
$F_F(4)$	Minimum critical angle across subtowers[†]
$F_D(\cdot)$	Exps. 2 and 4, In which direction?
$F_D(1)$	x,y angle of minimum critical angle
$F_D(2)$	x,y angle of minimum critical angle across subtowers
$F_B(\cdot)$	Exp. 5, Bump?
$F_B(1)$	Average distance from the table’s center
$F_B(2)$	Average x,y distance from the table’s center
$F_B(3)$	Average height
$F_B(4)$	Average distance from the nearest edge
$F_B(5)$	Average x,y distance from the nearest edge
$F_B(6)$	Minimum x,y distance from the nearest edge + height
$F_B(7)$	Minimum x,y distance from the nearest edge
$F_B(8)$	Like 1, except with maximum, instead of average
$F_B(9)$	Like 2, except with maximum, instead of average
$F_B(10)$	Like 3, except with maximum, instead of average
$F_B(11)$	Like 4, except with minimum, instead of average
$F_B(12)$	Like 5, except with minimum, instead of average
$F_B(13)$	Like 6, except with minimum, instead of average
$F_B(14)$	Like 7, except with minimum, instead of average

[*] Critical angle is defined as the angle of center-of-mass of the tower about the nearest edge (in the horizontal plane) of the convex hull around the tower’s base. Negative critical angles mean the center-of-mass is outside the convex hull (more unstable), and positive values mean it is inside. For F_B, each property was computed over the red blocks and over all of the blocks, and their ratio was used as the corresponding feature’s value.

[†] Subtower is defined as a disjoint (noncontacting) subset of the blocks in a tower. Multiple subtowers existed in some stimuli as a natural consequence of the random procedure by which towers were generated; these subtowers were supported by the ground, but were not in contact with each other.