Probabilistic simulation predicts human judgments about substance dynamics

James R. Kubricht*

Yixin Zhu*

Chenfanfu Jiang*

Demetri Terzopoulos

Song-Chun Zhu

Hongjing Lu

University of California, Los Angeles

Word count: 12487

Running head: Reasoning About the Dynamics of Substances

Address for correspondence:

James Kubricht
Department of Psychology
University of California, Los Angeles
405 Hilgard Ave.
Los Angeles, CA 90095
kubricht@ucla.edu

Abstract

The physical behavior of moving substances is highly complex, yet people can interact with them in their everyday lives with ease and proficiency. To investigate how humans achieve this remarkable ability, the present study examined human performance on an extension of the classical water-pouring problem (Schwartz & Black, 1999) and a substance dynamics prediction task adapted from previous work (Bates, Yildirim, Tenenbaum, & Battaglia, 2015). Participants were asked to perform three distinctively different tasks: (1) judging the relative pouring angle of two substance-filled containers which varied in the volume and viscosity of their contents; (2) predicting the resting geometry of sand pouring from a funnel onto a surface; and (3) predicting the dynamics of three substances—liquid, sand, and a collection of rigid balls—flowing past obstacles into two basins. Our findings indicate that people do not rely on simple qualitative heuristics based on physical attributes (i.e., viscosity, friction, and ball restitution) or perceptual variables (i.e., substance volume and position) when forming judgments and predictions. Instead, computational results from an intuitive substance engine (ISE) model employing probabilistic simulation support the hypothesis that humans infer future states of perceived physical situations by propagating noisy representations forward in time using approximated rational physics. The ISE model outperforms ground-truth physical models in each experiment, as well as competing non-simulation models based on data-driven learning approaches. Our results expand on previous work proposing human use of mental simulation in physical reasoning and demonstrate human proficiency in predicting the dynamics of physical events involving non-solid substances.

Keywords: Intuitive physics; mental simulation; animation; reasoning; substance representation; prediction

1. Introduction

Imagine that you are a server at a restaurant carrying dishes and beverages from a kitchen to customers' tables. To do this, you must decide where to hold the dishes and at what orientation to prevent their contents from spilling. More impressively, you must achieve this while navigating through tables, chairs, and customers in the environment. In the unfortunate case that a substance-filled container (e.g., a glass of water or a bowl of soup) topples over and spills onto an occupied table, you must also decide where the substance will travel and how it will interact with obstacles resting on the surface, hoping to intercept the fluid before it pours onto an unfortunate customer. We encounter similar situations frequently in our daily lives while interacting with non-solid substances ranging from granular materials (e.g., sugar, salt, or sand) to viscous liquids (e.g., syrup or honey), contained in receptacles of various shapes and sizes. How is the human cognitive system able to rapidly form predictions and judgments about the physical dynamics of substances to allow for such interactions?

1.1. Background

Over the past several decades, the field of intuitive physics has examined the human capacity to perceive and reason about situations in the physical world. These studies have primarily explored predictions and judgments that people make about rigid bodies—e.g., the path of a moving projectile or the relative weight of colliding objects (see Kubricht, Holyoak, & Lu, 2017 for a review)—rather than non-solid substances. Although comparatively fewer studies have explored human reasoning about the dynamics of substances, their results have given rise to quite perplexing results. For example, the Piagetian water-level-task (WLT; Rebelsky, 1964) was originally designed to determine at what developmental period children begin to represent and attend to horizontal referents in Euclidean space. In the task, an empty 2D container is rotated and displayed

on a piece of paper, and participants are instructed to draw the surface of the contained liquid, given that it intersects a specified location on the container's inside surface (see **Figure 1A**). The correct response is to draw a horizontal line which lies parallel to the bottom edge of the paper, but approximately 40% of *adults* draw water lines that deviate from the horizontal by 5 degrees or more (McAfee & Proffitt, 1991). This result appears to suggest that humans do not understand that liquid surfaces should remain horizontal regardless of the orientation of their containers, although this finding has generally been attributed to the mental representation of substance position relative to a rotated frame of reference fixed to the container (see McAfee & Proffitt, 1991). In other words, since there are no spatial referents near the container images to gauge rotation (e.g., a line indicating the ground or a vertical wall), the container edges themselves must be used as the "coordinate axes" for the represented situation. This interpretation is reinforced by research showing that employees in professions where containers are interacted with regularly (e.g., waitresses and bartenders) succumb to more errors on the WLT than employees working container-free jobs (Hecht & Proffitt, 1995).

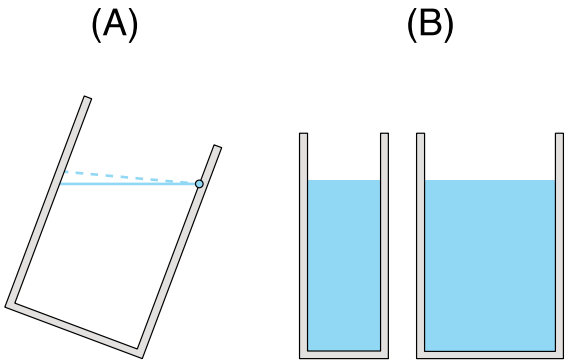

Figure 1. Stimulus images from (A) the water-level task (WLT) and (B) the water-pouring problem (WPP). (A) The solid line indicates the correct response where the surface of the contained water is horizontal; the dashed line indicates an incorrect response where the liquid's surface is rotated 5° from the correct position, in the direction of container rotation. (B) In the

WPP, the correct response is that the thinner container should spill last if rotated to the same degree as the wider container.

It is also likely that the impoverished format in which the problem was presented (i.e., on paper at a single time point) prevented participants from utilizing their intuitive knowledge about the physical behavior of substances in the real world, leading to their erroneous predictions. While the status of physical systems in the natural environment can be inferred and recognized from rich (frequently updated) visual inputs, explicit tasks utilizing static imagery convey comparatively little information. Indeed, research in intuitive physics has demonstrated inconsistencies between human performance on explicit and implicit reasoning tasks (Kaiser, Proffitt, Whelan, & Hecht, 1992; Kozhevnikov & Hegarty, 2001; Krist H. , 2000; Krist, Fieberg, & Wilkening, 1993; Smith, Battaglia, & Vul, 2013). As shown in many studies, people succumb to systematic errors when explaining a physical situation using idiosyncratic descriptive knowledge given impoverished visual information (e.g., drawing the status of a static situation across time) but can form accurate predictions and judgments about situations that are depicted dynamically via rich 3D visual information (Battaglia, Hamrick, & Tenenbaum, 2013; Kaiser, Proffitt, Whelan, & Hecht, 1992; Kubricht, Holyoak, & Lu, 2017; Smith, Battaglia, & Vul, 2013; Ye, et al., 2017). Thus, the WLT does not appear to trigger the commonsense knowledge people have about constraints in the physical world, such as early emerging sensitivity to core physical principles (see Baillargeon R. , 2004; Hespos, Ferry, Anderson, Hollenbeck, & Rips, 2016; Spelke, Katz, Purcell, Ehrlich, & Breinlinger, 1994).

Another example of the discrepancy between explicit and implicit performance in intuitive physics comes from the water-pouring problem (WPP; Schwartz & Black, 1999). This is a modification to the WLT that includes two containers—one wider than the other—filled to the same height with water (see **Figure 1B**). Participants solving the WPP must determine which

container needs to be tilted farther before the water inside begins to pour out. Surprisingly, only 34% of the participants in the study (averaged across container-type) correctly reported that a thinner container would need to be tilted farther than a wider one. However, when instructed to complete the task by closing their eyes and imagining the same situation, nearly all (95% of) participants rotated a thinner container filled with imaginary liquid farther. These findings demonstrate that people can reason successfully about the physical behavior of substances by mentally simulating an imaginary event, even if their corresponding explicit knowledge is inaccurate. They also show that people are more likely to utilize mental simulation when those systems are encountered in a realistic (dynamic) context instead of an ambiguous (static) one.

Taken together, the studies above demonstrate that people are capable of reasoning about the physical status of observed situations—including those involving non-solid substances—but fail to do so when task presentation is poor. In these cases, people appear to construct domain-specific physical theories (Cook & Breedin, 1994) that are inconsistent with their implicit expectations in the real world. Importantly, these erroneous theories can be diminished or even overcome when the problems are made less ambiguous (Kaiser, Jonides, & Alexander, 1986; Kaiser, Proffitt, Whelan, & Hecht, 1992).

A question that naturally arises is how problems can be framed to facilitate mental simulation and probabilistic inference in novel physical domains. The present study aims to address this question by focusing on judgments and predictions about physical events at critical moments, such as when a container will begin to spill, or where a moving substance will come to rest. Recent neural evidence suggests that people utilize an internal "physics engine" encoded within the brain's multiple-demand system to reason about physical situations via mental simulation (Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016). These events are represented to

encode both observable properties (e.g., position and volume) and hidden attributes (e.g., viscosity and friction) to enable physical inference (Hegarty, 2004). Hence, the role of dynamic context is particularly important because it provides visual motion cues from which substance attributes can be inferred (Kawabe, Maruya, Fleming, & Nishida, 2015). These attributes have been shown to influence people's predictions about the dynamics of substances. For example, participants solving the WPP rotated imaginary containers filled with molasses—a liquid with a relatively high viscosity attribute—farther than ones filled with water (Schwartz & Black, 1999). It is therefore essential to provide observers with rich visual information about substance behavior to facilitate attribute inference and subsequent mental simulation, particularly when the substance is unfamiliar.

The present paper takes the general approach of providing adequate information for people to estimate observable and hidden properties of substances in physical scenes and then examines whether these estimates enable predictions about future event states via mental simulation. In the next sections, we outline the goals and contributions of the study, suggest how implicit physical reasoning can be framed at the computational level, and propose a unified framework for performing this computation.

1.2. Goals and Contributions of the Study

The present study aims to determine whether an approximated simulation model coupled with noisy input variables can account for human predictions and judgments in a range of novel situations involving non-solid substance dynamics. In Experiment 1, participants were provided visual input that enabled the perception of physical variables (e.g., viscosity), and then were asked to perform a task by reasoning about the relative pouring angle of two containers filled with liquids differing in their volume and viscosity. In Experiment 2, participants reasoned about the dynamics

of sand (a granular material less ubiquitous in daily life than viscous liquids), and in Experiment 3, participants reasoned about the dynamics of liquid, sand, and rigid balls in a prediction task adapted from Bates et al.'s (2015) liquid experiments.

Our empirical paradigms extended Schwartz and Black's (1999) water-pouring problem and Bates et al.'s (2015) basin prediction problem, and incorporated critical features identified from previous physical reasoning literature (e.g., Kaiser, Proffitt & Anderson, 1985; Schwartz & Black, 1996; Smith, Battaglia & Vul, 2018). Our experiments provided observers with animated demonstrations of flow behavior that could guide inferences about unobservable attributes and probed the involvement of mental simulation processes in subsequent reasoning tasks. The experiments used dynamic and realistic displays to study reasoning mechanisms underpinning intuitive physics in a wide variety of scenarios involving a representative set of substance types (liquid, sand, and rigid objects). The experiments reported here included two critical features in their design and procedure: (1) Situations are presented in a dynamic context (at least once) to guide inferences about latent attributes and observable volumes/positions of physical entities. For example, videos demonstrating the movement of each substance are shown prior to each prediction and judgment task. (2) The task in each problem does not involve a description or explanation of the situation *across* time. For example, rather than asking participants to trace out the motion trajectory of a moving object, participants are asked to catch the object or predict where the object will land at a particular time point. These design characteristics provide general guidance to prompt the use of mental simulation for physical predictions and judgments.

The computational work presented in the paper provides a unified simulation method that is applied to situations involving the three substances (liquid, sand, and rigid objects), rather than developing separate simulation methods for each substance type. Under this approach, physical

models describing the behavior of each substance are formed from different parameterizations of the simulation. This would be analogous to humans utilizing a unified physics engine to handle arbitrary substances, where various physical transition models are formed depending on situational (perceptual) cues. Thus, the current modeling implementation is the first to examine predictions about multiple substances, ranging from liquids to rigid objects, while ensuring comparable comparisons of model predictions with human judgments. Although we do not suggest that the current simulation method is the only viable account of human mental simulation at the process level, its adaptability to different substance types indicates that separate simulation frameworks are not necessarily required to account for human judgments involving the dynamics of non-solid substances as well as rigid objects. As the mental simulation is a central component in physical reasoning, it is important to investigate human proficiency across diverse situations involving different substances, some more familiar than others. Although previous work has involved both rigid objects and non-solid substances (Bates, Yildirim, Tenenbaum, & Battaglia, 2015; Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Sanborn, 2014; Sanborn, Mansinghka, & Griffiths, 2013; Smith, Battaglia, & Vul, 2013), the current experiments extend findings to a wider variety of scenarios involving a representative set of substance types (liquid, sand, and rigid objects). Moreover, we explicitly examine people's predictions about the dynamics of granular substances (sand; see Experiments 2 and 3). This substance type is particularly interesting because it serves as a "middle ground" of sorts between liquid and rigid objects. The current study is—to our knowledge—the first to utilize a computational model based on physics-based simulation to explain human predictions about the dynamics of granular substances.

Finally, we address the plausibility of approximations to mental simulations for physical events. Physics-based simulations for non-rigid substances are computationally intensive because the partial differential equations which describe fluid and granular substance's behavior have no analytic solution and therefore must be integrated numerically. While modern physics-based simulators can perform this integration, it requires an immense amount of computation and time. In contrast, people appear to have little difficulty making predictions about each type of substances in different physical situations. For example, regardless of whether a container full of marbles or liquid spills onto a table, an average person will quickly (and implicitly) form a prediction about where the material will be in the future. How could humans achieve such reasoning mastery despite the large range of variations in physical complexity? In the current study, we explore the possibility that people represent non-solid substances as discrete collections of rigid balls (on the order of ten to thirty) and generate predictions by applying emulated principles of rigid-body mechanics to spatially represented variables. We term this significantly simplified physical model the *ball approximation* (BA) model and test its performance in comparison with human judgments. This modeling endeavor addresses the important question of whether complex laws of physical dynamics are necessary to account for people's predictions about substance behavior in novel situations, or whether the dynamics of liquids can be emulated using a low-resolution, particle-based representation. While particle-based approaches have shown success in previous work (see Bates et al., 2015), the current study extends the approach to a novel problem domain involving liquids varying in their viscosity and different substances. This extension is necessary because it clarifies the viability of rough physical approximations and helps to answer questions regarding their role in reasoning through mental simulation.

2. Intuitive Substance Engine: Background and Overview

A growing body of evidence suggests that intuitive physics is based on Bayesian inference over structured knowledge of physical principles and noisy perceptual inputs (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Sanborn, 2014; Sanborn, Mansinghka, & Griffiths, 2013). The problem of inferring physical properties (h) can be modeled as assessing the posterior probability of a candidate hypothesis ($h = H$) based on observable information O, and can be computed using Bayes' rule:

$$P(h = H|O) = \frac{P(O|h = H)P(h = H)}{\sum_{H'} P(O|h = H')P(h = H')} \ . \quad (1)$$

To enable the inference in Equation (1), a computational model needs to define the likelihood term $P(O|h)$ for evaluating the probability of observing the input data given certain physical properties, and the prior term $P(h)$ based on general knowledge of how physical properties are distributed in the physical world.

2.1 The Noisy Newton Framework

The *noisy Newton* framework for physical reasoning assumes that inferences about complex, dynamical systems can be generated by combining noisy perceptual inputs with the principles of classical (i.e., Newtonian) mechanics $P(O|h)$, given prior beliefs about physical and perceptual variables $P(h)$ (Bates, Yildirim, Tenenbaum, & Battaglia, 2015; Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Sanborn, 2014; Sanborn, Mansinghka, & Griffiths, 2013; Smith, Battaglia, & Vul, 2013). Under this framework, the locations, motions and physical attributes of objects and substances are sampled from distributions with physical and perceptual noise and propagated forwards in time using physics-based simulation models that approximate Newtonian mechanics. The status of the situation

throughout the simulation is then queried, and the outputs of the query are averaged across numerous simulations to determine the probability of the associated human judgment. The model has successfully explained a variety of human judgments across diverse physical situations, such as object collisions (Sanborn, 2014; Sanborn, Mansinghka, & Griffiths, 2013), block towers (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016), containment situations (Liang, Zhao, Zhu, & Zhu, 2015), and projectile motion (Smith, Battaglia, & Vul, 2013).

For simple physical events such as head-on object collisions, the likelihood term in Equation (1) can be calculated from four observable variables—the velocity of each object before and after impact—to form inferences based on candidate hypotheses $P(h)$. For each candidate hypothesis (e.g., prior beliefs about the attributes of each object), values for the corresponding observable variables can be determined analytically via the principle of conservation of momentum. Then, a likelihood value is obtained for each hypothesis, $h = H$, by comparing expected observations with ground-truth evidence from the real world using a noisy perception model. Thus, the noisy Newton model for object collisions predicts human judgments by answering three questions: (1) What are people's expectations about the physical characteristics of a dynamic scene; (2) What would the scene look like given those characteristics; and (3) How likely are the observable data given those expectations?

The noisy Newton model has also achieved success in situations involving towers of stacked (rectangular) blocks (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016). It has been used to predict human judgments about whether a block tower will fall down and in which direction (Battaglia, Hamrick, & Tenenbaum, 2013), as well as whether blocks of one color are heavier/lighter than blocks of another color (Hamrick,

Battaglia, Griffiths, & Tenenbaum, 2016). Unlike Sanborn et al.'s (2013) model for object collisions, the block tower model does not assess the likelihood of candidate hypotheses by comparing the expected velocity of each rigid body (i.e., each block) to observed data. After all, it is unlikely that people attend to *every* object's motion in a scene when perceiving and reasoning about their characteristics and future states. Instead, hypothesized scene states are queried throughout time to determine *whether* or *how* an event occurred, which is compared with the ground-truth outcome to assess likelihood. The key supposition here is that people possess a runnable mental model of rigid-body dynamics to simulate (noisily) perceived situations forwards in time.

Recent research has shown that the noisy Newton framework can be extended to explain people's predictions about the dynamics of liquid (Bates, Yildirim, Tenenbaum, & Battaglia, 2015). In their task, participants reasoned about which of two basins the majority of liquid would fall into after pouring past randomly generated 'obstacle courses.' To explain human performance, Bates et al. (2015) proposed an *intuitive fluid engine* (IFE), where future liquid states are approximated by probabilistic simulation via a Smoothed Particle Hydrodynamics method (SPH; Monaghan, 1992). The SPH method serves as an approximation to ground-truth physics, which is a key extension for applying the noisy Newton framework to liquid dynamics. The method approximates a volume of liquid as a set of particles with perceptual uncertainty drawn from a 2D Gaussian distribution. The attributes of each particle are updated by comparing them with the attributes of nearby 'neighbors' (closer neighbors have a larger impact). The new attributes are then used to update each particle's position and velocity at each iteration, and the process is repeated at each timestamp for every particle. Overall, Bates et al.'s (2015) model matched human

judgments about future liquid states and provided a better quantitative fit than alternative models that did not employ physical simulation or uncertainty about physical variables.

Bates et al.'s (2015) results point to an extraordinary conclusion: People are able to mentally simulate substance-related events with ease and proficiency, even though the physical equations governing the fluid motion (1) lack analytic solutions, (2) are particularly difficult to numerically integrate, and (3) require coarse approximation methods, such as SPH, to achieve correspondence with the real world. Furthermore, the researchers found that the learning-based model (i.e., a heuristic and deep learning account) failed to achieve comparable performance with simulation-based models for fluid dynamics. The current experiments follow a similar design, where ground-truth and probabilistic simulation results are compared with learning-based (data-driven) model performance. However, it remains unclear whether the success of Bates et al.'s (2015) model can extend to novel situations involving different types of human judgments and predictions. Moreover, the human capacity for reasoning about less common substances (e.g., granular materials, like sand) remains to be explored and modeled via a probabilistic simulation approach. Motivated by Bates et al.'s (2015) success, the current work aims to address these questions using a range of physical reasoning tasks, as well as a physics engine based on state-of-the-art simulation methods in computer graphics.

2.2. Intuitive Substance Engine (ISE)

The present study developed the same general class of model as Bates et al.'s (2015) IFE, which we term the *intuitive substance engine* (ISE). The ISE models physical predictions and judgments by simulating substance states forwards in time and querying perceptual/physical variables at critical time steps. Substance states are represented by the perceptual and physical variables that define their physical behavior, such as the position of each substance element and the physical

attributes (e.g., viscosity, density, pressure, etc.) which govern how those positions change over time. The state of a substance at time step t is denoted by S_t, where $t = 0, 1, \ldots, T$. Given an initial ground-truth substance state, $\bar{S}_0$ (i.e., the true values for each perceptual and physical variable prior to movement), the ISE first forms the distribution of an observed initial state S_0, $P(S_0|\bar{S}_0)$, reflecting noisy perception and prior beliefs about underlying variables.

The goal of the ISE model is to form expectations about human predictions and judgments in various intuitive physics tasks. To do this, the model must infer a *final* state distribution given *observed* states, $P(S_T|S_0, \bar{S}_0)$. This is achieved by sampling from the observed state distribution $i = 1, \ldots, N$ times and propagating each sampled observed state, $S_0^{(i)}$, forwards in time using the Material Point Method (MPM) simulation method, $M: S_t \rightarrow S_{t+1}$. We denote the state distribution of the entire sequence from $t = 0$ to $t = T$ as $P(S_{0:T}|S_{0:T-1}, \bar{S}_0)$. This distribution is then queried to form predicted response distributions, given different initial (ground-truth) substance states, $\bar{S}_0$. A graphical depiction of the ISE modules is shown in **Figure 2**, further information about the query functions used in each experiment are provided in **Appendix A**, and additional technical details on MPM is provided in **Appendix B**.

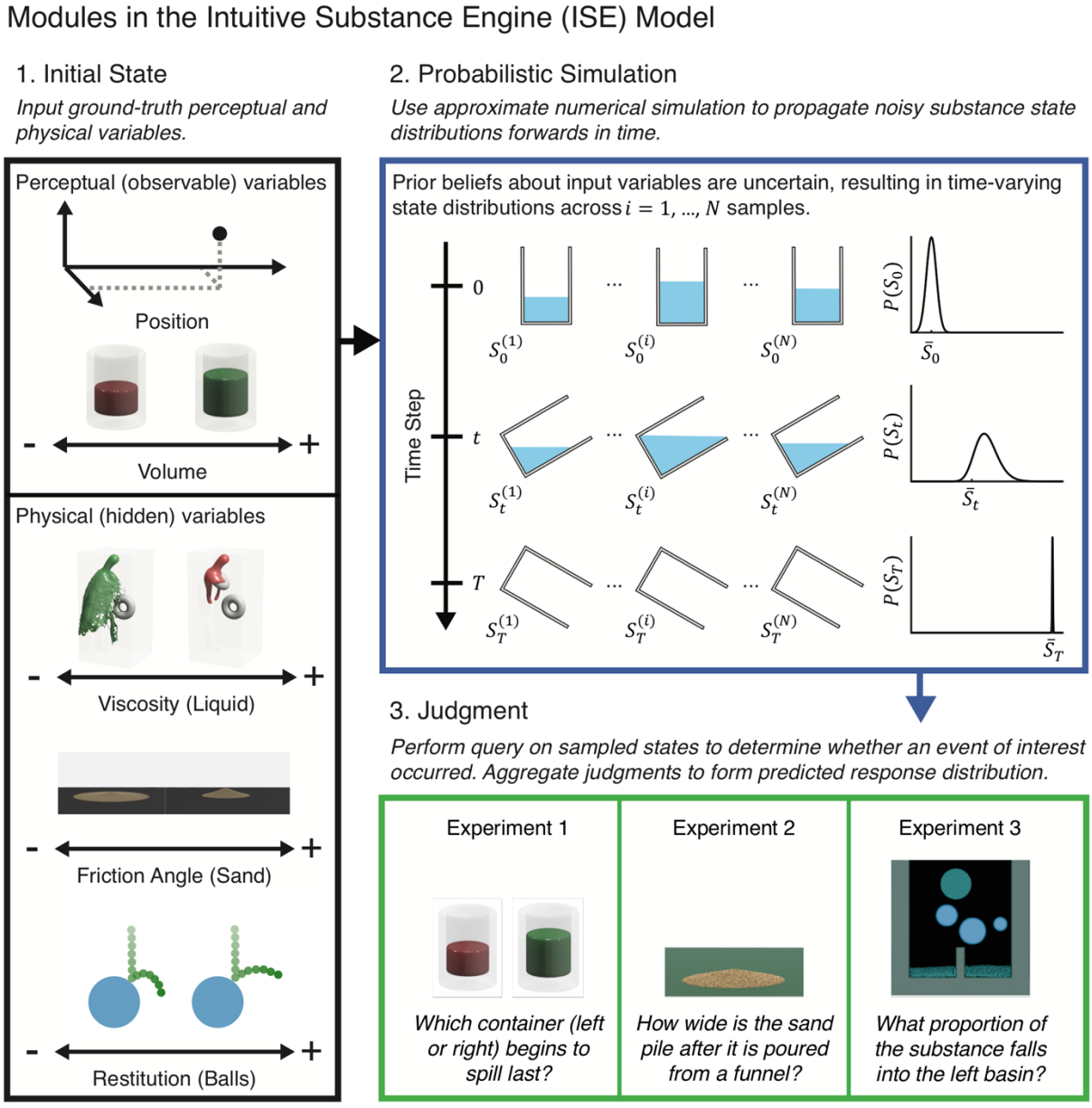

Figure 2. The three core modules of the ISE model are shown. (1) Input variables are separated into perceptual (observable) and physical (hidden) variables. Prior distributions are placed on the listed variables; other perceptual and physical variables are also passed to the simulator (e.g., flow velocity, density, etc.) but only their ground-truth values were used. (2) Substance states follow noisy distributions due to uncertain prior expectations about underlying variables. The ISE model uses the sampling approach with MPM simulation to derive the substance state distribution at each time step: $P(S_{0:T})$. (3) This distribution is queried based on the question asked to participants in each experiment, and queries are aggregated across $n = 1, \ldots, N$ samples to form predicted response distributions.

Although both our ISE model and Bates et al.'s (2015) IFE model are formed under the same computational framework, each employs different simulation methods as the physical simulation engine. The simulation of incompressible flows through numerical evaluation of physical equations has become one of the most significant topics in computer graphics and mechanical engineering. The velocity field of simulated fluids is determined according to the constraints specified in the Navier-Stokes equations. These partial differential equations place constraints on key physical properties (i.e., momentum and compressibility) which are quantified by underlying variables, such as localized substance velocity, density, pressure, and viscosity (see details **Appendix B**). To numerically solve these equations, our ISE model adopts the Material Point Method (MPM; Zhu & Bridson, 2005; Jiang C. , Schroeder, Selle, Teran, & Stomakhin, 2015), which has become the standard in physics-based simulation calculations due to its accuracy, stability, and efficiency. In recent work, a detailed evaluation of physical simulation methods in terms of perceived realism showed that the MPM method yielded highest realism scores compared with other methods for simulating fluid behavior (Um, Hu, & Thuerey, 2017). However, we do not make the claim that people use a particle/grid representation (as developed in MPM simulation method for physical simulation). Instead, the MPM method is utilized for two primary reasons. First, it provides more realistic visualizations of demonstration events, which facilitate human inferences about latent substance attributes. Second, the MPM method can be applied to each of the substances investigated in the current work, allowing for a unified modeling framework applicable to any substance type.

2.3. Uncertainty in Perceptual and Physical Variables

Fluid simulation with physical dynamics provides deterministic fluid movements if the ground-truth values of substance attributes, position, and volume are known. Hence, the decisions directly

derived from the MPM simulator are binary judgments, which implies that physical simulation with high precision cannot explain humans' probabilistic judgments in intuitive physics tasks. As demonstrated in the work by Bates et al. (2015) and the noisy Newton model (e.g., Sanborn et al., 2013), uncertainty in perceptual and physical variables plays an important role in accounting for people's physical judgments. In our ISE model, we combine the MPM simulator with noisy input variables (i.e., position, volume, viscosity, friction angle, and restitution), thereby accounting for physical uncertainty and the influence of perceptual and physical variables on the prediction and judgment tasks. The distributions used to generate noise—along with their parameters—are provided in the ISE Model Details section in each experiment.

2.4. Simulation-Based Approximation Model

To examine the possibility that intuitive physics can be achieved through low-resolution spatial representations and physical simulations, we further developed a model to coarsely approximate ground-truth physics and the behavior of non-solid substances. This approximation hypothesis is motivated by Bates et al.'s (2015) results, where researchers used a small number of particles (as few as 15) in their SPH simulation method to achieve high correlations with human predictions in estimating water movement patterns. To examine whether this approximation strategy may work across a range of substances, we constructed a simulation-based model which approximates substances as a set of rigid balls which interact with one another according to the principle of conservation of momentum: i.e., the same principle governing the dynamics of colliding balls. In the *ball approximation* (BA) model, latent attributes of liquid, such as viscosity, were approximated by damping the angular acceleration of each ball with a magnitude proportional to its angular velocity and weighted by a stiffness parameter. Our implementation of angular damping is akin to creating an imaginary lever attached to each ball, where the lever connects via a "ball-

and-socket" joint. The levers—and therefore each of the balls—are acted upon with a resistive force which is proportional to how fast each lever is rotating. Thus, a ball that rotates quickly with respect to the bottom of the container will experience a stronger resistive force than a ball with comparatively slower rotation. This approximation roughly emulates the dynamical impact of viscosity, which represents the internal friction in non-solid substances that prevents local deformation.

The approximation model can significantly reduce computation cost compared with the ISE model utilizing high-resolution, physics-based simulation. In the ISE model, each small particle represents a chunk of continuum material. A typical simulation employs at least 10k particles, with each particle containing (1) degree of freedoms of position, velocity, rotation, shearing, dilation; and (2) material parameters governing its dynamics: density, Young's modulus, bulk modulus, and internal friction. However, in the ball-approximation (BA) model, the material is approximated with around 30 rigid spheres in contrasts to tens thousands of particles in ISE model. The ratios for computational complexity, degree-of-freedom count, and the dimension of the dynamics space are thus approximately more than 1000/3 between the ISE and BA models.

3. Experiment 1: Reasoning about the relative pouring angle of liquid-filled containers

Experiment 1 aims to demonstrate that people can utilize mental simulation to form judgments about the dynamics of liquids which vary in their viscosity attribute. The experimental task is designed to conform with the three features outlined in Section 1.2 to facilitate spatial representation and subsequent mental simulation. To quantify the extent that people employ inferred attributes of non-solid substances when reasoning about novel situations, we utilized a recent development in graphical substance simulation (Bridson, 2015; Jiang C. , Schroeder, Selle,

Teran, & Stomakhin, 2015) to capture the dynamic behavior of non-solid materials in vivid animations. Previous work has shown that realistic animations can facilitate representation of *dynamic* physical situations (Tversky, Morrison, & Betrancourt, 2002). Furthermore, recent research on human visual recognition indicates that latent attributes of liquids (e.g., viscosity) are primarily perceived from visual motion cues (Kawabe, Maruya, Fleming, & Nishida, 2015). Therefore, displaying realistic substance behavior is important to perceive the key physical attributes that facilitate mental simulation.

3.1. Participants

A total of 152 participants (99 females; mean age = 20.7 years) were recruited from the Department of Psychology subject pool at the University of California, Los Angeles, and were compensated with course credit.

3.2. Materials and Procedure

Prior to the reasoning task, participants viewed animated demonstrations of the movement of a moderately viscous liquid in two situations. The liquid used in the demonstrations was colored orange and was not observed in the judgment task. In the first (flow) demonstration, the orange liquid pours over two torus-shaped obstructions in a video looped three times and lasting for 11.5 seconds. The flow demonstration videos were presented to provide visual motion cues to inform participants' inferred viscosity values[1]. Following the flow demonstration, participants viewed a video of a cylindrical container filled with the same orange liquid tilting at a constant angular rate ($\omega = 22\,° \cdot \sec^{-1}$) from the upright orientation of the container and moving towards the horizontal. The tilting demonstration video was looped three times for a duration of 14.7 seconds[2].

[1] The flow demonstration video can be viewed at https://vimeo.com/339876565/2cdf12885b
[2] The tilting demonstration video can be viewed at https://vimeo.com/339876553/600e5180cc

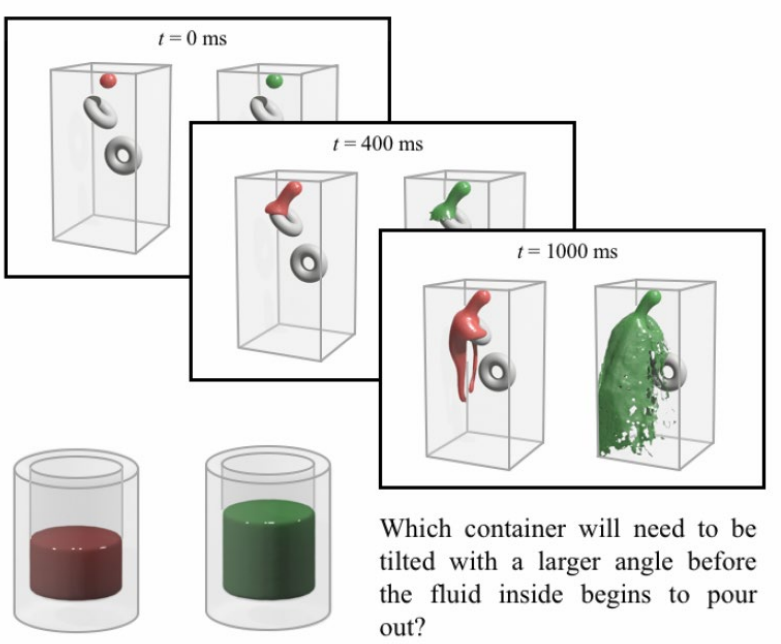

Figure 3. Illustration of flow demonstration video and judgment trial. (Top) Sample frames from the high viscosity fluid (*Hvisc*, red) and low viscosity fluid (*Lvisc*, green) demonstration videos. (Bottom) Tilt judgment trial, where the proportion of each container filled with the *Hvisc* and *Lvisc* liquid is 40% and 60%, respectively. Participants were asked to report which container would need to be tilted with a larger angle before the liquid inside begins to pour out.

Following the demonstration videos, two new liquids were introduced, one with low viscosity (*Lvisc*; similar to water) and one with high viscosity (*Hvisc*; similar to a thin syrup). The *Lvisc* and *Hvisc* liquids were colored either red or green, and the color was counter-balanced across participants. As shown in the top panel of **Figure 3**, participants viewed a flow demonstration video of both the *Hvisc* and *Lvisc* liquids (looped three times) for a duration of 11.5 seconds before each judgment trial. The two flow videos were presented side by side for comparison, and the relative position of each liquid was counterbalanced across participants. The *Lvisc* and *Hvisc* liquids were selected to readily distinguish each one based on their perceived viscosities, which were inferred from visual motion cues in the flow demonstration videos (see Kawabe, Maruya, Fleming, & Nishida, 2015)[3].

[3] The stimulus videos can be viewed at https://vimeo.com/339876567/59c9266c7d

In the subsequent reasoning task, participants viewed a static image of two containers side by side filled with the *Lvisc* and *Hvisc* liquids (see bottom panel of **Figure 3**). Participants were instructed to assume that each container was tilted simultaneously in the same way as observed earlier for the orange liquid in the tilting demonstration. They were informed that both containers were tilted at the same rate and were provided with the quantity of liquid in each container. Participants were then asked to report which container would need to be tilted with a larger angle before the liquid inside begins to pour out and received no feedback following the completion of each trial. The container images remained on the screen until a response was made; there was no time limit for making a response on each trial. The experiment manipulated the volume of the *Lvisc* and *Hvisc* liquids (V_L and V_H, respectively) in each container across the values 20%, 40%, 60%, and 80%, representing the proportion of the container filled. Hence, the experiment consisted of 16 trials presented in a randomized order, including all possible volume pairs between the *Lvisc* and *Hvisc* liquids. The experiment lasted approximately 10 minutes.

3.3. Human Results

The proportion of participants choosing the *Hvisc* liquid container as pouring last with a larger titled angle for each judgment trial is shown in the top-left panel of **Figure 4**. To assess the relationship between *Hvisc* liquid volume and human judgments, a repeated-measures ANOVA was conducted across two within-subjects factors: i.e., *Lvisc* and *Hvisc* liquid volume with four levels each. The two-way interaction between *Lvisc* and *Hvisc* liquid volume was significant, $F(9, 143) = 45.12$, $p < .001$, $\eta_p^2 = 0.74$, indicating that the effect of *Hvisc* liquid volume on *Hvisc* response proportion varied according to the quantity of *Lvisc* liquid in the alternative container. Since the two liquid volume variables interacted, data were separated into each *Lvisc* liquid volume condition and, the main effect of *Hvisc* liquid volume was examined. Results indicate that *Hvisc*

liquid volume was least influential on participants' judgments when the *Lvisc* container was 80% filled, $F(3, 149) = 2.36$, $p = .07$, $\eta_p^2 = 0.05$. However, for smaller *Lvisc* volumes, *Hvisc* liquid volume had a significant impact on the pouring judgment. For example, *Hvisc* liquid volume showed a significant main effect when the *Lvisc* container was filled to a lesser extent: $F(3, 149) = 70.77$, $p < .001$, $\eta_p^2 = 0.59$ for V_L = 60%; $F(3, 149) = 143.12$, $p < .001$, $\eta_p^2 = 0.74$ for V_L = 40%; $F(3, 149) = 107.22$, $p < .001$, $\eta_p^2 = 0.68$ for V_L = 20%. In other words, when the *Lvisc* liquid container had the greatest volume, participants consistently reported that the *Hvisc* container would pour last regardless of how much *Hvisc* liquid was in the other container. However, participants were increasingly hesitant to report that the *Hvisc* liquid container would pour last when the *Lvisc* liquid container was filled to lesser volumes. These results indicate that participants attended to both liquid volume and viscosity when forming their relative pour angle judgments.

3.4. Can Heuristic-Based Reasoning Account for Human Performance?

Next, we examined whether people rely on heuristic-based reasoning to make their judgments. One candidate heuristic is that given two containers filled with different volumes of each liquid, the container with lesser liquid volume requires a greater rotation before beginning to pour. While participants adhered to this rule for trials where the volume of *Hvisc* liquid (V_H) was less than *Lvisc* liquid volume (V_L) (i.e., $V_H < V_L$), their judgments for each of the $V_L < V_H$ trials did not accord to the same heuristic. For example, in trials where volume difference was least salient (i.e., $V_H = 40\%, 60\%, \text{and } 80\%$ and $V_L = 20\%, 40\%, \text{and } 60\%$, respectively), the lesser-volume heuristic predicts *Lvisc* liquid responses. However, *Hvisc* response proportions for those trials were significantly greater than zero, $t(151) = 9.92, 8.86, 8.10$, $p < .001$, $\text{Cohen's } d = .80, .72, .66, \text{respectively}$. A second potential heuristic is to always choose the *Hvisc* liquid as requiring a greater rotation based on general knowledge that *Hvisc* liquid moves slower than *Lvisc*

liquid. The above three cases also disagreed with this heuristic since *Hvisc* response proportions were significantly less than one, $t(151)$ = 15.22, 17.04, 18.65, $p < .001$, Cohen's d = 1.23, 1.38, 1.51, respectively. In summary, response proportions in the specified trials reveal that participants attended to latent liquid attributes (e.g., viscosity) *and* volume difference when making their tilt angle judgments.

It is worth noting that one possibility is that participants reasoned by applying multiple heuristics via a combination rule. For instance, the *Hvisc* container is chosen if both liquid volumes are equal, and the container with the least liquid volume is chosen otherwise. Although this combined heuristic could provide a qualitative account to the trends of human responses, it fails to predict the quantitative performance across the conditions. Furthermore, results from a follow-up study utilizing realistic depictions of water and honey provides results again this heuristic combination hypothesis. Specifically, participants consistently responded that the *Hvisc* liquid (honey) would pour last, even in trials where the *Lvisc* liquid (water) volume was less than that of the *Hvisc* liquid (honey). These results suggest that participants were unlikely to utilize heuristic-based reasoning in the 10-minute experiment without any explicit feedback, although it is possible that participants might develop and apply such rules as their familiarity with the liquid-pouring environment increases through experience.

3.5. Simulation-Based ISE and BA Model Details

As described in Section 2.2 (see **Appendix A** and **B** for additional technical details), we adopt an MPM-based simulation method for our ISE model. The physics-based MPM simulator was used to determine the ground-truth about which container would need to be tilted with a larger angle before the liquid inside begins to pour out.

The input variables for the ISE in Experiment 1 were liquid volume with perceptual uncertainty and liquid viscosity with physical uncertainty. Given the ground-truth value of volume ($V_L^{(GT)}, V_H^{(GT)}$), and viscosity for each liquid ($\mu_L^{(GT)}, \mu_H^{(GT)}$), N = 10,000 noisy samples ($\{(V_L^{(i)}, V_H^{(i)}, \mu_L^{(i)}, \mu_H^{(i)}),\ i = 1, \dots, \mathrm{N}\}$) were generated and passed to the MPM simulator. The simulator propagated each sampled situation forwards in time and determined when each container's contents began to spill over the rim of the cylinder. The container which required a longer duration to spill was chosen as the predicted response for each sample. By aggregating predictions across the 10,000 samples, the ISE outputs a predicted response distribution for each trial.

To model perceptual and physical uncertainty in participants' mental simulations, the ISE sampled liquid volumes and viscosities from noisy distributions reflecting imperfect volume estimation and viscosity inference via the visual system. Gaussian noise (0 mean, σ_V^2 variance) was added to the ground-truth Lvisc and *Hvisc* volume, $V_L^{(GT)}$ and $V_H^{(GT)}$. Gaussian noise (0 mean, σ_μ^2 variance) was also added to a scaled viscosity value for each liquid, $c_L \cdot \mu_L^{(GT)}$ and $c_H \cdot \mu_H^{(GT)}$. The scale variable c was a free parameter in the ISE model and was included to account for participants' biased estimates of viscosity for each liquid. For example, if people have a prior belief that liquids should behave like water, then their inferred viscosity estimates would be negatively biased: i.e., $c < 1$. Viscosity uncertainty was added to these scaled viscosity values in logarithmic space (see Sanborn, Mansinghka, & Griffiths, 2013): $\mu_i = \log^{-1}(\log(c \cdot \mu^{(GT)}) + \varepsilon)$, where $c \cdot \mu^{(GT)}$ is the scaled viscosity value and ε represents Gaussian noise with 0 mean and σ_μ^2 variance. Logarithmic noise corresponds with increasing uncertainty for larger variable values. The results reported herein used the following model parameters: $c_L = 2.5$, $c_H = 1.1$, $\sigma_V = 0.15$, and $\sigma_\mu = 0.15$. Note that the current reported parameter $c_L = 2.5$ seems to be large, but it in fact

is very small with respect to the common viscosity values; for instance (with unit mPa·s), water (1.0016), milk (2.12), oil (56.2), honey (2000-10000), ketchup (5000-20000).

We also ran simulations with the ball-approximation (BA) model to examine whether an approximation model with a low-resolution of spatial representations and physical simulations can account for human judgments. The BA simulations used the noisy viscosity as in the ISE model, i.e., noisy stiffness was generated by offsetting a mean value parameter with Gaussian noise on a logarithmic scale. The material is approximated using 30 rigid spheres. The results reported herein used the following model parameters: $c_L = 2.1$, $c_H = 1.6$, $\sigma_V = 0.1$, and $\sigma_\mu = 0.17$.

3.6. Non-Simulation Model Details

To examine whether substance simulation is necessary to account for how humans reason about liquid dynamics, we compare the ISE model with two statistical learning methods: the generalized linear model (GLM; McCullagh, 1984) and eXtreme Gradient Boosting (XGBoost; Chen & Guestrin, 2016). These models are purely data-driven, meaning that they learn from examples and do not encode any explicit knowledge of physical laws or physical simulation. The selected features for these models include: (1) the volume of liquid in each container; and (2) the viscosity value of the *Lvisc* and *Hvisc* liquids.

To predict the human judgment for the i^*th trial, denoted as J_{i*}, both non-simulation models were trained with 15 trials $\{J_i, i = 1, 2, \ldots, 16, i \neq i^*\}$ and tested with the i^*th trial. Specifically, each training trail was augmented with noisy input and generated 10,000 samples. Since GLM is capable of making a prediction in continuous variable space, the trained GLM model is directly applied to the test trial to predict how likely a container will need to be tilted to a larger angle before the liquid inside begins to pour out. Since the XGBoost provides a (direct)

discriminative classification (i.e., +1 indicating selection of the left container and -1 indicating selection of the right container), we introduced noise (the same method for the ISE) to the input variables (volume and viscosity features) for each test trial. For each test trial, a set with 10,000 samples was generated. The trained XGBoost model is applied to classify the labels (+1 or -1) in each sample, which are then aggregated to form the predicted response proportion for each test trial.

3.7. Model Comparisons

We first compared how well different computational models account for human performance for the 16 judgment trials. The top right panels of Figure 4 show the predictions from the ground-truth model, which yielded a low correlation with human judgments, $r(14) = 0.066$. The low correlation is due to the ground-truth model generating binary decisions across conditions, due to the absence of any perceptual/physical noise. Therefore, the model failed to account for the gradual change of human performance as a function of liquid volume.

The remaining panels in **Figure 4** depict results from the ISE, BA, GLM, and XGBoost models with perceptual/physical noise. Although human judgments and model predictions were highly correlated as $r(14) = 0.995\ (ISE), 0.97(BA), 0.95(GLM),$ and 0.93(XGBoost), the ISE model showed the greatest correlation with human performance. Root-mean-squared deviation (RMSD) between human judgments and the models' predictions were 0.0989, 0.0946, 0.13, and 0.12, respectively, which showed the smaller deviation for simulation-based models (ISE and BA). In comparison to the data-driven models (i.e., the GLM and XGBoost models), the simulation-based ISE and BA models utilize real and approximated material properties (viscosity/stiffness) *and* perceptual features (volume) as variables in a generative physical model. This computational approach provides a better account to human predictions in the current physical reasoning task.

Table 1 shows the summary of the model comparison results. In summary, the ISE model predictions were most correlated with human judgment. We examined model performance using the Bayesian information criterion (BIC) to account for the different number of free parameters in each model. We found that the BA model showed the best fitting result with the lowest BIC value. These results support the role of physical simulation as a potential mental model to account for human performance in intuitive physics tasks. Furthermore, the success of the BA model demonstrates that precise numerical simulation methods are not necessarily required to provide a good fit to human performance; instead, an approximation with reduced computation efforts is adequate to account for human performance through mental simulation. The worse performance from data-driven models (XGBoost and GLM) shows the inadequacy of the generic learning-from-data approach, and highlights the importance of making inference according to the laws of physics or the approximated forms in reasoning tasks related to physical events.

Table 1. Comparison between human performance and model predictions for the ground-truth, ISE, XGBoost, and GLM models in Experiment 1. The root-mean-squared-deviation (RMSD, lower value indicates better model fitting result) and correlation are shown, in addition to the number of free parameters in each model and the corresponding Bayesian information criterion (BIC) score (lower value indicates superior model fit). The bold text indicates the best model performance according to the different performance measures.

	Ground-Truth Model	ISE	BA	XGBoost	GLM
# params	5	9	5	8	7
Correlation	0.066	**0.995**	0.97	0.93	0.95
RMSD	0.67	**0.0989**	0.0998	0.13	0.12
BIC	1.05	-49.08	**-59.88**	-43.11	-48.44

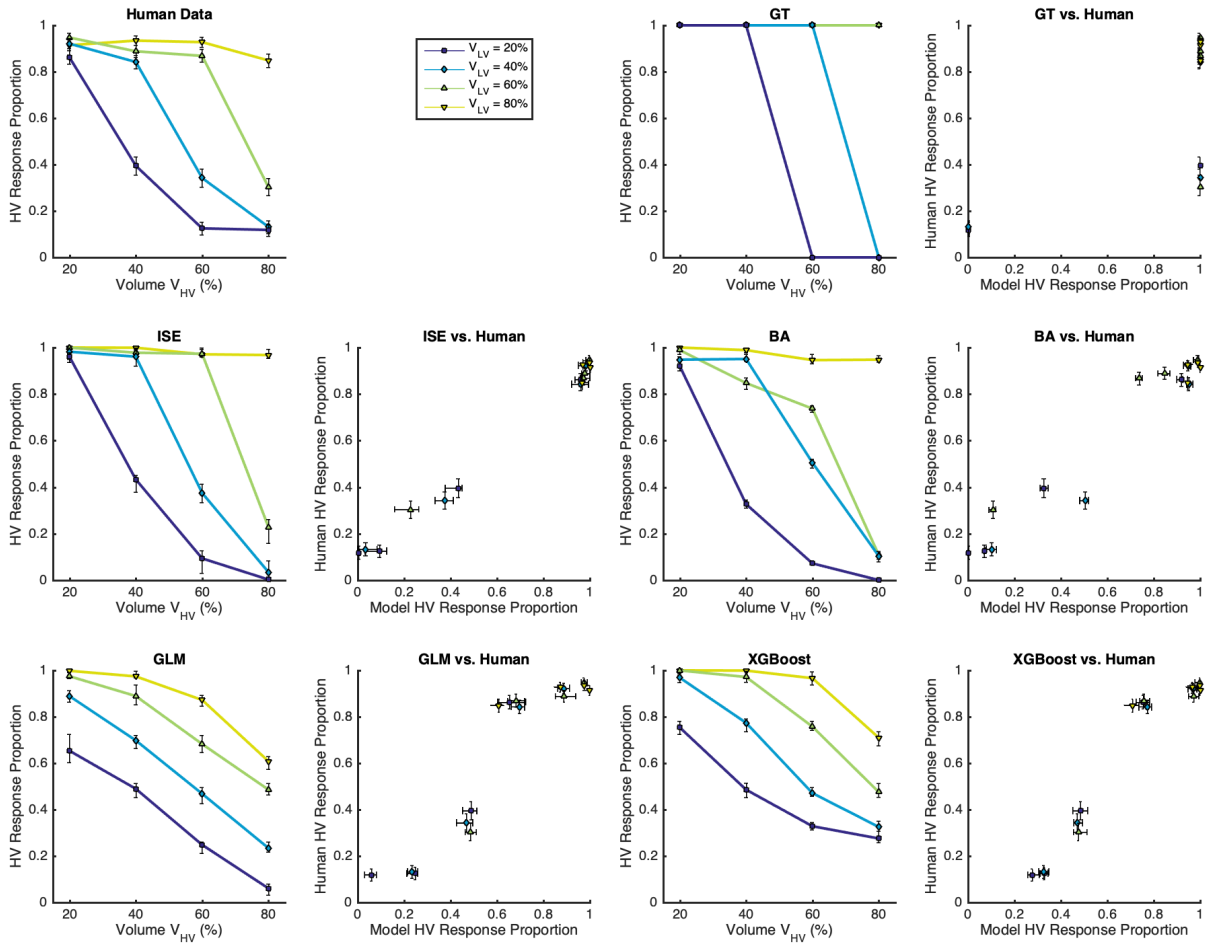

Figure 4. Human response proportions and predictions from the five candidate models: (Top-Right) Ground-Truth (GT) Model, (Middle-Left) Intuitive Substance Engine (ISE) Model, (Middle-Right) Ball Approximation (BA) model, (Bottom-Left) General Linear Model (GLM), and (Bottom-Right) eXtreme Gradient Boosting (XGBoost) Model. Columns 1 and 3 provide human judgments and model predictions; horizontal axes indicate *Hvisc* liquid volume, and vertical axes indicate the proportion of *Hvisc* liquid responses associated with a greater rotation angle. Columns 2 and 4 compare model predictions with human judgments; vertical lines indicate the CIs of human judgments, and horizontal lines indicate the CIs of model predictions. The ISE and BA simulation models outperform competing non-simulation data-driven models.

4. Experiment 2: Reasoning about Granular Material

Results from Experiment 1 indicate that people are sensitive to the viscosity attribute of liquids when reasoning about the dynamics of non-solid substances, and the simulation-based models provides the best account for human performance in a range of conditions. However, it remains

unclear whether this proficiency extends to substances that are less ubiquitous in daily life than liquid—specifically granular materials such as sand, and whether the simulation-based models still provides a good account for human performance. The second experiment was designed to determine whether humans can predict the resting geometry of a volume of sand after it is poured from a funnel onto a surface, and whether dynamic visualizations of the pouring behavior facilitate mental simulation of sand-surface interactions.

4.1. Participants

A total of 43 undergraduate students (31 females; mean age = 20.2 years) were recruited from the University of California, Los Angeles (UCLA), Department of Psychology subject pool and were compensated with course credit.

4.2. Materials and Procedure

Participants first viewed a simulated demonstration video of sand falling from a funnel suspended 10 cm above a level surface to form the final resting pile. The pouring event was viewed three times from a zoomed-out perspective (**Figure 5A**) and then a zoomed-in perspective (**Figure 5B**) for a duration of 35 sec[4].

After viewing the demonstration video, participants were shown a sand-filled funnel suspended 0.5, 1, 2, and 4 cm above the surface in a randomized order. After viewing each situation in both a zoomed-out (**Figure 5A**) and zoomed-in view (**Figure 5B**), participants were asked to indicate which of four sand piles would result from the sand pouring from the funnel at the indicated height (**Figure 5C**). For each trial, the stimulus images remained on the screen until a response was made. The pile choices were shown from the zoomed-in perspective and represented the ground-truth resting geometries resulting from each situation, which were generated by the

[4] The demonstration video can be viewed at https://vimeo.com/339881783/42fc4e5589

MPM physical simulator with ground-truth initial conditions (funnel height, friction angle, sand volume, etc.). According to the ground-truth simulations, Piles #1, #2, #3, and #4 correspond with the piles resulting from sand pouring from funnels suspended 0.5, 1, 2, and 4 cm above the surface, respectively. Therefore, the correct choice for sand pouring from a 0.5 cm funnel height was Pile #1, the correct choice for sand pouring from a 1 cm funnel height was Pile #2, etc. The experiment consisted of 4 trials presented in a randomized order.

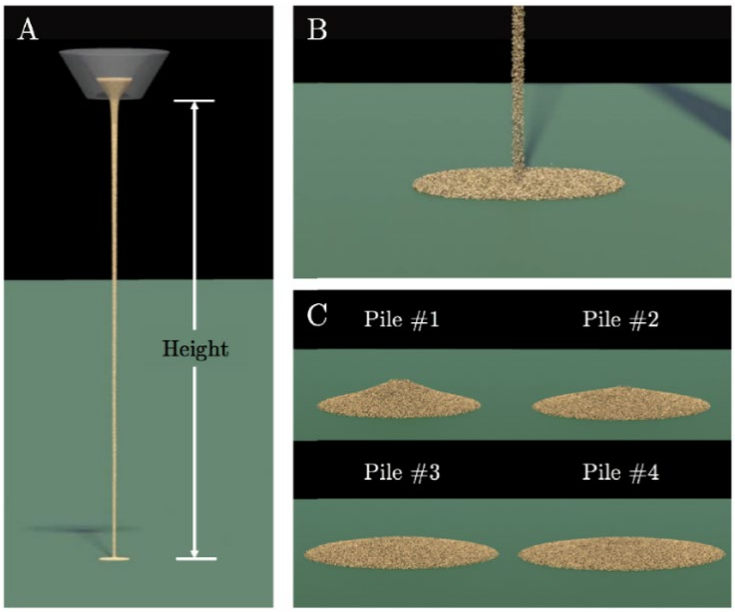

Figure 5. Intermediate frames from the demonstration video in Experiment 2 from the (A) zoomed-out and (B) zoomed-in view. (C) Sand pile choices in Experiment 2's judgment task are displayed.

4.3. Human Results

Figure 6 (left) shows participant's response proportions as a function of funnel height, as Trial #1 (h = 0.5 cm), Trial #2 (h = 1 cm), Trial #3 (h = 2 cm), Trial #4 (h = 4 cm). Participants' pile choices varied significantly across different funnel heights, $\chi^2(9) = 176.54$, $p < .001$, Cramer's $V = 0.74$, as pile choices shifted towards higher-numbered, flatter piles as funnel height increased. This finding suggests that humans take into consideration of the sand's falling distance to influence their predictions about the sand's resting geometry.

However, in comparison with the ground-truth simulation, the impact of funnel height on the judgment of the resulting sand pile revealed the bias shift in human judgements. With the small funnel height in Trial #1 (h = 0.5 cm) and Trial #2 (h = 1 cm), although the ground-truth model predicts that participants should choose Pile #1 in Trial #1, Pile #2 in Trial #2, human responses revealed a tendency towards choosing lower-numbered, more conical piles. This bias continues as funnel height increases. For example, most participants incorrectly chose Pile #2 in Trial #3 (h = 2 cm), and the proportion choosing Piles #2 and #3 were also higher in Trial #4 (h = 4 cm). Hence, in comparison with the ground-truth model, human performance did not show strong correlation with ground-truth predictions, $r(12) = 0.17$. These results indicate that participants' predictions were sensitive to funnel height, but humans showed clear tendency of perceiving more conical resting states compared with the ground-truth simulation.

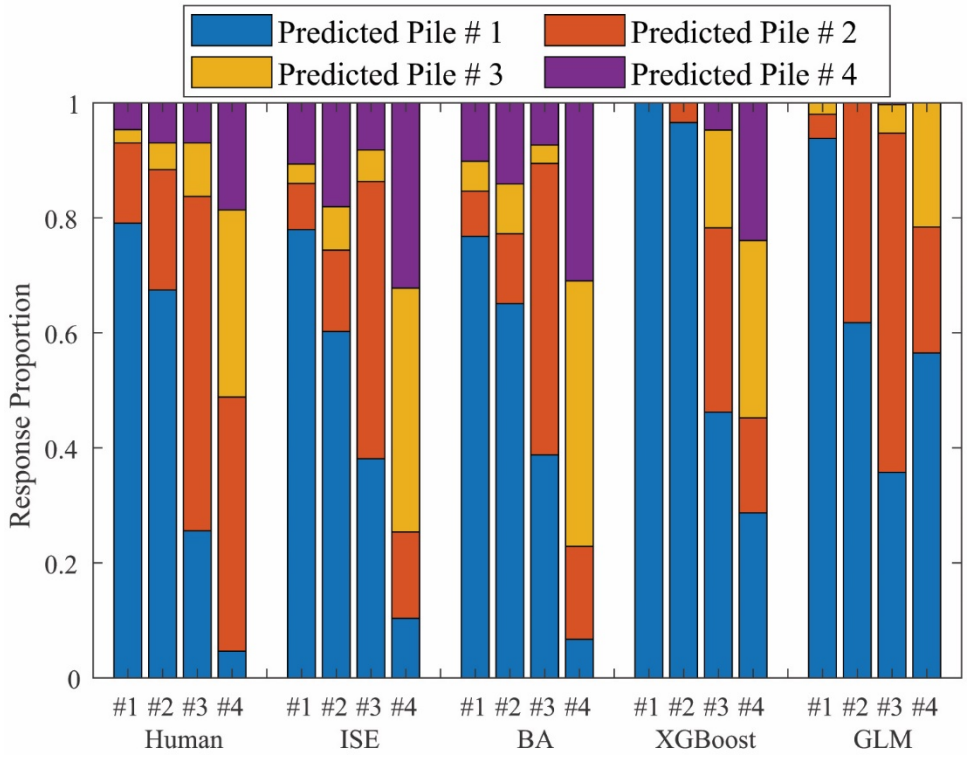

Figure 6. Model prediction results compared to human judgments. The five groups (left to right) correspond to human judgment, ISE, BA, XGBoost, and GLM. Each bar (#1, #2, #3, and #4) corresponds to testing trials with funnel height 0.5, 1, 2, and 4 cm, respectively. The ground-truth

predictions are Pile #1 choices for Trial #1 (blue bar), Pile #2 choices for Trial #2 (magenta bar), Pile #3 choices for Trial #3 (yellow bar), and Pile #4 choices for Trial #4 (purple bar). Participants' pile choices were consistently biased towards lower-numbered (flatter) piles.

4.4. Simulation-Based ISE and BA Model Details

The input variables for the ISE in Experiment 2 were funnel height (i.e., initial sand height) with perceptual uncertainty and sand friction angle with noise. Given the ground-truth values of initial funnel height and friction angle ($H^{(GT)}, \theta^{(GT)}$), N = 10,000 noisy samples ($\{(H^{(i)}, \theta^{(i)}), \mathrm{i} = 1, \ldots, \mathrm{N}\}$) were generated and passed to the MPM simulator, which returned the final height of the sand pile for each sample. Instead of choosing from 4 piles (i.e., the task presented to the participants), the MPM simulator compares the estimated height of the final sand pile, formally $D(H^{(i)}, \theta^{(i)}) = H_p \in \mathbb{R} > 0$, with the heights of the 4 pile options given to participants. The pile option with the minimum height difference was chosen as the predicted judgment for each sample. Finally, by aggregating predictions across the 10,000 samples, the ISE outputs a predicted response distribution for each trial.

To model uncertainty in participants' mental simulations, the ISE sampled funnel heights and friction angles from noisy distributions. Gaussian noise (0 mean, σ_H^2 variance) was added to the ground-truth funnel height in each situation. Gaussian noise was also added to the ground-truth friction angle $\theta^{(GT)}$ in logarithmic space (see Sanborn, Mansinghka, & Griffiths, 2013): $\theta^{(i)} = f^{-1}(f(\theta_i^{(GT)}) + \varepsilon)$, where $\theta^{(GT)}$ is the ground truth value of the initial sand height, $f(\theta_i^{(GT)}) = \log(\omega|\theta_i^{(GT)}|)$, and ε represents Gaussian noise with 0 mean and σ_ε^2 variance. Note that unlike in Experiment 1 where liquids were defined through demonstration, Experiment 2 shows substances with a known label (such as sand). Hence, the ground-truth attribute values were not scaled in Experiment 2. This was due to the simulated sands employing physical attribute values that are

consistent with sands encountered regularly in daily life. The results reported herein used the following model parameters: $\sigma_H = 0.11, \sigma_\varepsilon = 0.65, \text{and } \omega = 0.85$.

We further explore if the BA model – combining low-resolution spatial representations and physical simulations – can properly account for human judgments. The BA model approximates the simulation of the ISE model but with less parameters due to the reduced simulation complexity. The results reported herein used the following model parameters: $\sigma_H = 0.17, \sigma_\varepsilon = 0.42, \text{and } \omega = 0.58$. The material is approximated using 45 rigid spheres.

4.5. Non-Simulation Model Details

To examine the crucial role of mental simulation in physical reasoning about sands, two non-simulation models, GLM and XGBoost, were used as baseline models. The two models were trained on the three piles and tested on the remaining i^*th pile ($i = 1, 2, 3, 4, i \neq i^*$). During training, 10,000 samples were drawn for each remaining pile (30,000 samples total) and passed to the MPM simulator. Each sample was generated by adding noise to both funnel heights and friction angles, in the same way as the method used for the simulation-based ISE model. After training on the 30,000 samples, both non-simulation models were tested on another 10,000 samples generated from noisy input based on the configuration of Pile i^*. The final distribution was formed by aggregating the predictions across the 10,000 samples.

4.6. Model Comparisons

Figure 6 depicts human judgments in the sand pile reasoning task and the predictions of the simulation-based models (ISE and BA), and two data-driven models (XGBoost, GLM models). All four models showed correlations with human performance, $r(12)$ = 0.91, 0.90, 0.86, and 0.77 for ISE, BA, XGBoost, and GLM, respectively. However, the ISE model predictions showed the best account to human judgements than did the competing data-driven model predictions. We also

calculated the root-mean-square deviation (RMSD) between human response proportions and model results to compare the model fits. RMSD between human responses and ISE predictions for the four judgment trials was less than that between all the other models (see **Table 2**). After considering the number of free parameters in different models, the BA model provided the best fit to the human data than other models. These results reveal that physical simulation serves as a potential mental model to provide the best account for human performance in the task.

Table 2. Comparison between human performance and model predictions for the ISE, BA, XGBoost, and GLM models in Experiment 2. High correlation value, low RMSD, and low BIC score indicate better fitting results. The bold text indicates the best model performance according to the different performance measures.

	Ground-truth model	ISE	BA	XGBoost	GLM
# params	5	8	4	7	6
correlation	0.169	**0.907**	0.899	0.859	0.768
RMSD	0.458	**0.101**	0.104	0.183	0.195
BIC	-11.13	-51.18	**-61.34**	-34.94	-35.68

5. Experiment 3: Reasoning about Complex Interactions between Substances and Rigid Obstacles

Our results from Experiment 2 indicate that people can predict the resting geometry of sand poured from a funnel, even though they may not have rich experience interacting with granular substances in daily-life. Experiment 3 was conducted to determine (1) whether humans can reason about complex interactions between substances and rigid obstacles; and (2) whether their predictions about the resting state of sand in novel situations differ from predictions about other more familiar substances, such as liquid and rigid balls.

5.1. Participants

A total of 90 undergraduate students (66 females; mean age 20.9), were recruited from the UCLA Department of Psychology subject pool and were compensated with course credit.

5.2. Materials and Procedure

The procedure in Experiment 3 was similar to the design in Bates et al.'s (2015) experiment: i.e., participants viewed a quantity of a substance suspended in the air above obstacles and were asked to predict the proportion that would fall into two basins separated by a vertical divider below (see ***Figure 7***). The present experiment differed from previous work in that participants reasoned about the resting state of one of three different substances: liquid, sand, or sets of rigid balls. Also, whereas the previous study used polygonal obstacles, those in the present study were circles varying in size. Depth information was also not present in the rendered situations.

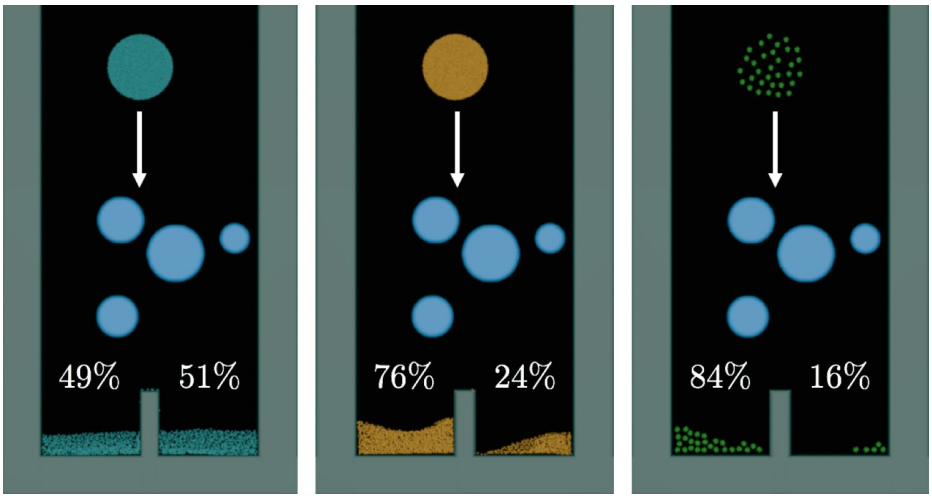

Figure 7. Initial (top) and final (bottom) state of liquid (left), sand (middle), and a set of rigid balls (right) for a testing trial in Experiment 3 with 4 obstacles. Number of obstacles varied between 2 and 5 in the testing trials. The percentages indicate the amount of each substance that fell into the left and right basins. Only the initial state of each substance was shown in the testing trials. The stimulus videos can be viewed at https://vimeo.com/339881779/81fdb529ef.

Situations were generated by sampling between 2 and 5 obstacle locations from a uniform distribution bounded by the width and height of the chamber. The diameter, d, of each obstacle

was sampled from a uniform distribution: $d \sim U(0.15, 0.85)$. The center points for each set of obstacles were generated by uniformly sampling the entire width and height of the chamber. If the generated obstacles were placed outside of the boundary, the configuration was rejected and re-sampled. For each substance, forty testing trials (10 trials with 2, 3, 4, and 5 obstacles) were chosen from the generated set such that the ground-truth proportion of *liquid* in the left basin was approximately uniform across trials: $L \sim U(0, 1)$. Importantly, the positions of the substance and configurations of obstacles were matched for each substance.

Participants were randomly assigned to either the liquid, sand, or rigid balls condition. Thirty participants were assigned to each condition in a between-subjects experimental design. Prior to the testing trials, participants completed five practice trials with two obstacles in each situation in a randomized order. After answering (1) which basin the majority of the substance would fall into and (2) the expected proportion that would fall into the indicated basin, participants viewed a demonstration video (13 second duration) of the situation unfolding and were told the resulting proportion in the ground-truth simulation. The videos were shown to provide participants with visual information to infer each substance's respective attribute: i.e., viscosity, friction angle, or restitution. After completing the practice trials, participants completed 40 testing trials in a randomized order by answering the same two questions in each trial. No feedback was given following the completion of each testing trial.

5.3. Human Results

The physics-based MPM simulator was used to determine the ground-truth proportion of each substance in the left and right basins for each of the generated situations. Participants' response proportions for the left-basin in the testing trials were correlated with ground-truth predictions in the liquid, sand, and rigid balls conditions, $r(38)$ = 0.86, 0.82, and 0.88; RMSD = 0.145, 0.170,

0.186, respectively, suggesting human judgments are overall consistent with physical models. However, we also found human judgments showed deviations from the ground-truth model predictions. We analyzed the deviation for each trial by subtracting the ground-truth proportion from each participant's proportion response. The deviation differed significantly between the three substance conditions, $F(2, 87) = 3.64$, $p = 0.03$, indicating that the difference between human predictions and the ground-truth status varied according to the substance type. The analysis further revealed that rigid balls predictions showed the largest deviation, and liquid predictions showed the least deviation. The next section examines whether the ISE and two non-simulation models can capture differences in human performance between the three substances.

The following analysis compares human basin predictions between experimental conditions. To determine whether participants' response proportions differed between substances, a repeated-measures ANOVA was conducted with one within-subjects factor (trial) and one between-subjects factor (substance type, or condition). We found that response proportions showed significant differences depending on substance type, $F(2, 87) = 5.72$, $p < 0.01$, $\eta_p^2 = .12$, indicating that participants accounted for substance properties when making their predictions through mental simulation. The reasoning for this claim is that the substances appeared similar to one another in each trial (i.e., all substance volumes had the same starting diameter), and only visual motion cues from the demonstration videos could be used to distinguish between substances. However, the rigid balls were somewhat dissimilar to the liquid and sand substances due to each ball being separated a small distance from its neighbors. Therefore, a second analysis was conducted which examined predictions for liquid and sand only. Again, response proportions varied significantly between these two substance types, $F(1, 58) = 11.84$, $p < 0.01$, $\eta_p^2 = .17$.

5.4. Simulation-Based ISE and BA Model Details

In the simulations for Experiment 3, the observable input variables for our ISE for each substance were (1) the initial, horizontal location of the substance; and (2) the locations of the circular obstacles in each situation. The latent substance attributes accepted by the engine were viscosity, friction angle, and restitution coefficient for liquid, sand, and the rigid balls, respectively. Given ground-truth values of substance position, obstacle position (2-5 obstacles), and substance attributes ($L_S^{(GT)}, L_O^{(GT)}, \mu^{(GT)}, \theta^{(GT)}, \varepsilon^{(GT)}$), N = 2000 samples (40 situations × 50 noisy samples) were generated ($\{\left(L_S^{(i)}, L_O^{(i)}, \mu^{(i)}, \theta^{(i)}, \varepsilon^{(i)}\right),\ i = 1, \ldots, \mathrm{N}\}$). Gaussian noise was added to the substance's (ground-truth) horizontal position (0 mean, σ_S^2 variance) and the obstacles' (ground-truth) positions in 2D space (0 mean, σ_O^2 variance). Logarithmic Gaussian noise was added to each substance's ground-truth attribute value via the logarithmic transformation specified in Experiment 2. The results reported here utilized the following model parameters for all three substances: σ_S = 0.59, σ_O = 0.63, σ_ε = 0.5, ω = 0.8. Recall that σ_ε is the Gaussian (perceptual) uncertainty in logarithmic space, and ω is the weight parameter in the log transformation. Two thousand samples (40 situations × 50 noisy samples) were drawn for each substance.

We further conducted simulations with the BA model – combining a low-resolution of spatial representations and physical simulations. The results reported herein used the following parameters σ_S = 0.38, σ_O = 0.53, σ_ε = 0.6, ω = 0.6. The material is approximated using 45 rigid spheres.

5.5. Non-Simulation Model Details

Similar to the previous experiments, two data-driven models GLM and XGBoost were tested. The training data were randomly generated situations with basin proportions calculated using resting state output from our MPM simulator. Input features were the collection of both the observable input variables and latent substance attributes used in the ISE prediction. In total, 6000 samples were used for training, and 2000 samples for testing.

5.6. Model Comparisons

Figure 8 depicts the comparison between human and model basin predictions from the ground-truth (GT), ISE, GLM, and XGBoost models. The ISE model predictions account for human judgment better than the ground-truth model predictions—$r(38)$ = 0.92, 0.92, 0.93; RMSD = 0.082, 0.085, 0.104 for liquid, sand, and rigid balls, respectively—indicating a superior account of human predictions across a range of substances. In comparison, the data-driven models GLM and XGBoost provided worse fit to human predictions, GLM: $r(38)$ = 0.77, 0.78, 0.64, RMSD = 0.078, 0.120, 0.193; XGBoost: $r(38)$ = 0.67, 0.74, 0.71, RMSD = 1.382, 1.422, 2.067 for liquid, sand, and rigid balls, respectively.

As in the previous experiment, we compared BIC measures of the models in each condition to account for the number of free parameters in each model. As shown in Table 3, we found that the BIC values of the ISE model were consistently less than the ground-truth, GLM, and XGBoost models for all three substances, liquid, sand, and rigid balls—further confirming the superior performance of the simulation-based ISE model than the data-driven model and the ground-truth physical model.

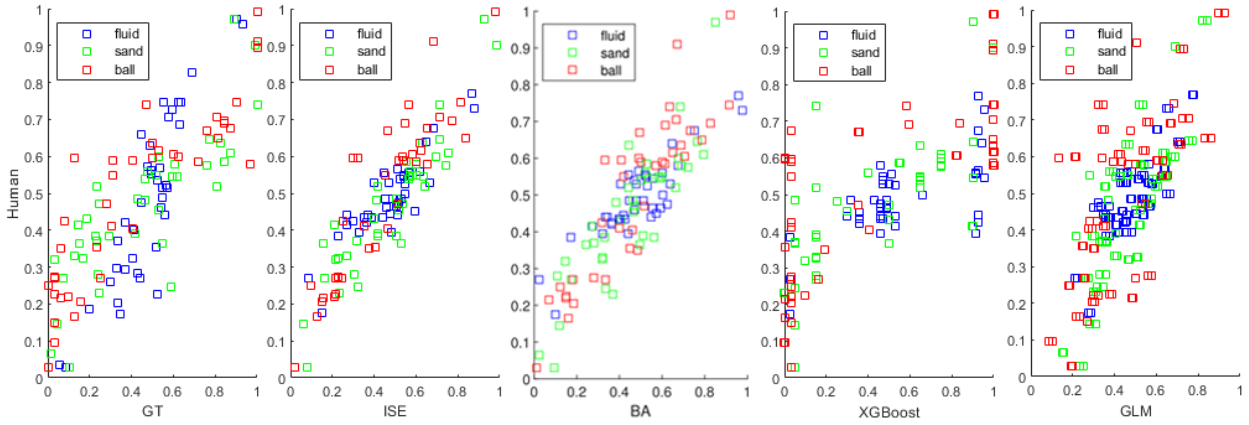

Figure 8. Model (left-basin) predictions compared with human predictions. The y-axis indicates participants' mean basin predictions. The x-axis depicts basin predictions from each model. From left to right: Ground-truth (GT), ISE, BA, XGBoost, and GLM. Separate colors indicate the type of substance.

Table 3. Model predictions and performance measures for the ISE, XGBoost, and GLM models in Experiment 3. The root-mean-squared-deviation (RMSD) is shown, in addition to the number of free parameters in each model and corresponding Bayesian information criterion (BIC) score (lower value indicates superior model fit).

	Ground-truth model	ISE	BA	XGBoost	GLM
# params (liquid, sand, rigid ball)	5 5 1	9 9 4	5 5 5	8 8 8	7 7 7
Correlation (liquid, sand, rigid ball)	0.86, 0.82, 0.88	**0.92,** **0.92,** **0.93**	0.88, 0.89, 0.91	0.67, 0.74, 0.71	0.77, 0.78, 0.64
RMSD (liquid, sand, rigid ball)	0.145 0.170, 0.186	**0.082,** **0.085,** **0.104**	0.103, 0.102, 0.111	1.382, 1.422, 2.067	0.078, 0.120, 0.193
BIC (liquid, sand, rigid ball)	-136.04, -123.31, -116.12	**-166.88,** -164.01, -147.87	-163.40, **-164.18,** **-157.41**	55.39, 57.68, 87.60	-178.26, -143.80, -105.78

It is worth noting that our ISE achieved consistent performance across all three substances, whereas GLM and XGBoost were less capable of predicting human judgments about rigid balls and liquid. In addition, the ISE model used only one third of the training samples that XGBoost and GLM needed, demonstrating that a generative physical model with noisy perceptual inputs is capable of learning with a smaller number of samples than data-driven methods.

6. General Discussion

Results from the experiments reported herein provide converging evidence that humans can predict outcomes of novel physical situations involving non-solid substances by propagating approximate spatial representations forwards in time using mental simulation. This stands in contrast to early research in rigid-body collisions suggesting that human physical predictions do not obey ground-truth physics, instead relying on heuristics (Gilden & Proffitt, 1994; Runeson, Juslin, & Olsson, 2000). Overall, our results agree with Bates et al.'s (2015) findings: i.e., ISE predictions entailing the noisy Newton framework outperformed ground-truth and data-driven models in each experiment, further reinforcing the critical role of perceptual noise and physical dynamics in intuitive physics reasoning.

Our results also indicate that people naturally attend to latent attributes when reasoning about familiar and unfamiliar substance states following observation of realistic demonstration animations. Although mental simulation has been demonstrated as a default strategy in other mechanical reasoning tasks (Hegarty, 2004; Clement, 1994), the participants in Schwartz and Black's (1999) experiments failed to spontaneously represent and simulate physical properties relevant to the judgment task. By designing tasks with regard to the three features outlined in Section 1, our participants were able to mentally simulate dynamic events and did not appear to rely on explicit or heuristic-based reasoning. While the present study indicates a set of simulation-

inducing task characteristics, further research should aim to determine specific experimental factors that trigger simulation strategies. Specifically, can the conditions employed in the present tasks extend to classical rigid-body and fluid dynamics problems to resolve the discrepancy between people's explicit predictions and tacit judgments, and if so, what additional task characteristics serve to facilitate mental simulation?

6.1. Are Precise Numerical Simulation Methods Needed to Explain Human Performance?

Taken together, our results demonstrate that human predictions and judgments about substance dynamics can be accounted for by a unified simulation method with uncertainty implemented into underlying physical variables. However, classical research in artificial intelligence has dismissed robust mental simulation as a strategy for physical reasoning due to its inherent complexity, often proposing simplified qualitative models instead (De Kleer & Brown, 1984). While the ISE model employed in the current study requires extensive numerical evaluation to make predictions about future substance states, humans appear to do so with precision and accuracy in comparatively small amounts of time. Results from the BA model provide evidence that humans approximate the dynamics of substances in a manner *consistent* with ground-truth physics based on coarse representations of substance elements. This circumvents one of the key problems in applying particle-based simulation methods to human physical inference: i.e., it is unlikely that people represent substances as collections of hundreds (or thousands) of constituent particles when mentally simulating their movements across time. Our BA model results further demonstrate that precise numerical evaluation of ground-truth differential equations is not needed to account for human judgments in our viscous liquid-pouring task. Success of the proposed physical approximation also reinforces the application of learning-based (deep learning) approaches, which encode how discrete objects *interact* with one another through observation of large sets of dynamic

training stimuli (Battaglia & Pascanu, 2016; Chang, Ullman, Torralba, & Tenenbaum, 2016; Grzeszczuk, Terzopoulos, & Hinton, 1998). Recent research has shown success in predicting the dynamics of deformable objects (Mrowca, et al., 2018) and non-solid substances (Li, Tedrake, Tenenbaum, & Torralba, 2018) by representing entities as collections of interacting particles. The success of our BA model, in addition to the cited work, suggests a possible algorithm-level explanation for how people may learn to emulate the physical laws underlying substance dynamics throughout long periods of observation in daily life. Importantly, underlying physical knowledge does not need to be "written into" these learning-based models for them to simulate the dynamics of entities in novel situations.

Another benefit of the BA model approach is that uncertainty can be distributed throughout the balls which comprise the substance "whole". When a substance is encountered in the real world, each of its portions likely entails varying magnitudes of uncertainty. For example, if a substance is spilled onto a table, an observer will likely be more uncertain about the positions and movements of substance elements that are moving in the center of the table compared with elements near the edge. After all, the risk that substance elements near the center will spill onto the observer is minimal. Furthermore, when a liquid is poured from a container, elements near the rim are most likely given more attention since their positions and motions indicate whether spilling is imminent. It is unclear, however, whether particle-based simulation methods will converge to a stable solution when substance element positions and motions are jittered with varying magnitude, especially if uncertainty is added *during* the simulated event. However, the positions and motions of rigid objects can be perturbed at any point during a simulated event without the risk of causing their solutions to diverge (Smith & Vul, 2013). Thus, the BA model affords the implementation of

multiple uncertainty modules, which can be used to further examine how situations are approximated over the course of mental simulation events.

6.2. Uncertainty in Numerical Physics Simulation

While human results are generally consistent with the physics-based simulation models coupled with noisy input variables, there remain discrepancies between model predictions and human judgments. Although our ISE and BA models accounted for perceptual uncertainty in each situation, the simulations themselves closely approximated normative physical principles. This assumption ignores a key component of uncertainty in mental simulation, namely that physical entities and their motions across time are determined via a precise generative model with high spatiotemporal resolution. In other words, given the same initial conditions, our ISE and BA models would arrive at the same outcome each time. Adding "stochastic noise" to physical dynamics (i.e., mental simulation uncertainty), however, has been shown to increase model performance when predicting human responses in simple physical situations (Smith & Vul, 2013). While mental simulation uncertainty can easily be built into rigid-body collisions—e.g., by perturbing the simulated position, speed, and/or direction of a projectile at various points in time—employing this strategy in the present physical simulations would preclude stable numerical evaluation. One potential avenue for introducing stochastic noise into numerical simulators (e.g., MPM, SPH, etc.) would be to add noise into the attributes themselves, which influence underlying substance dynamics. However, while it makes sense to suppose that the spatially represented positions or motions of discretized substance particles may change with time, it is unclear whether human estimates of the underlying attributes also fluctuate. We hope that future work utilizing substance simulation engines in intuitive physics will pursue this direction of research.

The most significant reason for exploring the role of stochastic uncertainty in physical inference via mental simulation is to take pressure off of perceptual uncertainty model components in accounting for biases in human predictions. Notably, if other components of uncertainty are at play, our current models might overestimate the magnitude of perceptual noise to account for their contribution. However, this question requires more than a purely computational approach. Specifically, empirical evidence supporting the role of perceptual and/or stochastic uncertainty in mental simulation is needed. Moreover, independent measurements of the magnitude of each type of uncertainty would allow for the comparison between fitted model parameters and those recorded directly from human cognition. It is the authors' opinion that the aforementioned pursuit would greatly support the noisy Newton framework as a cognitive model for human physical inference.

6.3. Developing Intuitions About Physical Dynamics

Although the computational results reported herein support the role of noisy and approximate mental simulation when reasoning about novel physical situations, it remains unclear how this capacity develops in humans. A breadth of findings in the developmental literature have investigated the periods of development in which sensitivity to core physical principles is established (Baillargeon R. , 1994; Baillargeon R. , 2004; Kotovsky & Baillargeon, 1998; Spelke, Breinlinger, Macomber, & Jacobsen, 1992). Across many physical phenomena (e.g., support, solidity, continuity, etc.), findings show that infants initially represent variables in a qualitative "all or none" fashion: e.g., when deciding whether a surface will support a box resting on top of it, infants first represent whether or not the box is completely on top of the surface. As the infant ages and gains experience in a given domain, she eventually learns to attend to the *proportion* of the box resting on the surface. In other words, primitive variable distinctions and corresponding rules which govern physical intuitions become more sophisticated over time. A core question is

how sophisticated to these rules and variables become? And if rules are enriched to the point where they closely approximate ground-truth physical principles and enable mental simulation, when does this capacity emerge?

Previous developmental literature suggests that infants initially utilize categorical (rule-based) reasoning when forming physical expectations (Baillargeon R. , 2002). For example, in the context of collision situations, infants as young as six months old understand that larger motor objects launch projectiles farther than smaller ones. However, it is also possible that infants are utilizing mental simulation based on highly simplified approximations to physical dynamics to form their expectations. The second hypothesis is consistent with recent developmental research showing that 12-month-old infants' expectations about moving objects are consistent with a probabilistic inference model employing abstract principles of object motion (Teglas, et al., 2011). It is tempting to propose that humans—regardless of their age—utilize one reasoning system over the other, or that one reasoning system preempts the other in development. However, it may be that physical intuitions rely on both reasoning systems, regardless of age. Moreover, these systems might engage with and enrich one another as experience with the world develops. In the context of the current experiments, we did not find evidence that heuristic reasoning was utilized to form predictions and judgments. However, it is possible that useful heuristics could be learned with experience in order to alleviate the cognitive resources needed for mental simulation. It is our hope that future research will investigate this interplay and individual differences in intuitive physics.

6.4. Concluding Remarks

Taken together, the current study provides evidence that people's predictions and judgments about the dynamics of substances are consistent with a computational model utilizing approximated Newtonian principles and noisy perceptual/physical inputs. The empirical findings

reveal that people are highly sensitive to the different dynamics of liquid, sand, and rigid objects in performing physical reasoning tasks. Finally, the modeling results show that a simulation method which approximates substances as a collection of rigid balls can provide a good account for human performance in lieu of advanced numerical simulation methods.

The results reported herein demonstrate that human expectations about the dynamics of non-solid substances are consistent with ground-truth physics, given that their representations are prone to observational and physical uncertainty. Moreover, people can predict the future status of situations involving granular materials (i.e., sand), which is less common in daily life than viscous liquids or collections of rigid objects. Results from the BA model also demonstrate the viability of coarse spatial approximation in substance-related situations, suggesting that human observers could be doing something similar when reasoning about the dynamics of liquids and other deformable materials.

Appendix A: Additional ISE Details

Comparison between MPM and SPH

The ISE model presented in the paper is formed under the same framework as Bates et al.'s (2015) IFE, although the two models use different physical simulation methods. Our ISE model employs the MPM simulation method, whereas the IFE model uses the SPH method. Indeed, SPH is a viable method for simulating the dynamics of liquids, granular materials, and colliding objects, although MPM provides a more efficient and accurate means of doing so (see Appendix B for further details). We do not envision that the predictions of the two methods would differ substantially from one another when applied to a given set of stimuli.

Query Functions

In the experiments reported herein, participants were tasked with reasoning about different physical properties of substances in dynamic scenes. The aspect of the physical environment critical to each task ranged from (1) when a liquid element exceeded a critical *position* in 3D space (Experiment 1); (2) the 3D *geometry* of a sand pile resting on a table (Experiment 2); and (3) the *volume* of liquid contained in two alternative receptacles, or basins (Experiment 3). The ISE model matches human expectations about different situations by probing different aspects of the respective environment. This is achieved by *querying* substance states (drawn from predictive probability distributions) and aggregating query results to form predicted response distributions. The query functions used in each experiment differ from one another, both in their variables of interest, as well as the point in time in which they are applied. We summarize the corresponding functions for each experiment as follows.

Experiment 1: When will it pour?

The query function Q_{pour} is applied to the entire state sequence $S_{0:T}$. The query function calculates the time step in which one of the two substances begins to spill over the rim of its respective container. The container whose substance which has not yet poured is output as the predicted choice.

Experiment 2: Which sand pile is correct?

The query function Q_{pile} is applied to the final substance state S_T. The query function calculates the diameter of the sand pile at the end of the simulation and is compared with the ground-truth diameter of each pile choice. The closest match is output as the predicted choice.

Experiment 3: Which basin will the substance fall into?

The query function Q_{basin} is applied to the final substance state S_T. The query function calculates the proportion of the substance which fell into the left basin. This proportion is output as the predicted estimate.

Predicted Judgment

Given the above query functions, we further define the predicted judgment for each trial in each experiment. The predicted judgment J_q for each query q is calculated by finding the expected value of each query function over the state sequence $S_{0:T}$:

$$J_q = E[Q_q|\bar{S}_0] \tag{2}$$

$$J_q = \int_{S_{0:T}} Q_q(S_{0:T})\, P(S_{0:T}|S_{0:T-1}, \bar{S}_0). \tag{3}$$

Appendix B: MPM Simulator Details

The Material Point Method (MPM) produces physically accurate and visually realistic simulations of the dynamics of liquid (Jiang C. , Schroeder, Teran, Stomakhin, & Selle, 2016) and sand (Klár,

et al., 2016), in addition to general continuum materials such as stiff elastic objects (Jiang C. , Schroeder, Teran, Stomakhin, & Selle, 2016). Unlike the Smoothed Particle Hydrodynamics (SPH) method—which purely relies on particles to discretize the computational domain—the MPM uses both particles and a background Eulerian grid. The Navier-Stokes equations are solved on the grid, allowing for: (1) accurate derivative calculations; (2) well-defined free surface and solid boundary conditions; and (3) an accurate first-order approximation of physical reality. The MPM also circumvents common artifacts of SPH, such as underestimated density near free surfaces and weakly compressible artifacts. In fact, the requirement for incompressibility is crucial in the substance dynamics problems studied in the present study. We chose not to use SPH because it does not guarantee a divergence-free velocity field unless additional computational components are included. MPM, however, maintains the benefits of particle-based methods due to its hybrid particle/grid nature. The presence of particles in the current model serves to facilitate visualization and the tracking of material properties. Besides modeling liquids, the state-of-the-art physics-based simulation methods have also provided realistic cues for modeling complex tool and tool-uses (Zhu, Zhao, & Zhu, 2015), generic containers (Liang, Zhao, Zhu, & Zhu, 2015), and soft human body dynamics (Zhu, Jiang, Zhao, Terzopoulos, & Zhu, 2016). The following paragraphs present a mathematical overview of our MPM simulator, which provides a unified, particle-based simulation framework that handles rigid balls, liquid, and sand with essentially the same numerical algorithm, albeit with appropriately differing material parameters. The MPM method is physically accurate, numerically stable, and computationally efficient, enabling us to synthesize a large set of stimuli in a short amount of time by simply varying material parameters and the locations of the initial objects and colliding geometries. Running each simulation in the same framework for the purposes of the present study also enables fair comparisons among the three types of substances,

since we avoid potential inconsistencies in the numerical accuracies of multiple simulators specialized to each material.

The differential equations governing particle-based substance dynamics utilize the principles of conservation of mass and momentum:

$$\frac{D\rho}{Dt} + \rho\nabla \cdot \boldsymbol{v} = 0, \tag{4}$$

$$\frac{D\boldsymbol{v}}{Dt} = \nabla \cdot \boldsymbol{\sigma} + \rho\boldsymbol{g}, \tag{5}$$

where ρ is the density of the simulated substance, $\boldsymbol{\sigma}$ is the stress imparted on a particle, $\boldsymbol{g}$ is gravitational acceleration, and $\frac{D}{Dt}$ is the material derivative with respect to time. The equations are discretized spatially and temporally with a collection of Lagrangian particles (or material points) and a background Eulerian grid. The material type of the simulated substances is naturally specified from the constitutive model, which defines how a material exerts internal stress (or forces) as a result of deformation.

Rigid balls are simulated as highly stiff elastic objects with the neo-Hookean hyperelasticity model, described through the elastic energy density function:

$$\Psi(\boldsymbol{F}) = \frac{\mu}{2}(tr(\boldsymbol{F}^T\boldsymbol{F}) - d) - \mu \log(J) + \frac{\lambda}{2}\log^2(J), \tag{6}$$

where d is the dimension (2 or 3), $\boldsymbol{F}$ is the deformation gradient (i.e., the gradient of the deformation from undeformed space to deformed space), J is the determinant of $\boldsymbol{F}$, and both μ and λ are Lamé parameters that describe the material's stiffness.

Liquid is modeled as a nearly incompressible fluid, with its state governed by the Tait equation (Batchelor, 2000):

$$p = k\left[\left(\frac{\rho_0}{\rho}\right)^{\gamma} - 1\right], \tag{7}$$

where p is the pressure, ρ and ρ_0 are the current and original densities of the particles, $\gamma = 7$ for water, and k is the bulk modulus (i.e., how incompressible the liquid is). Through this Equation-of-State (EOS), the stress inside a non-viscous liquid is given by $\sigma = -p\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We further adopt the Affine Particle-In-Cell method (APIC; Jiang C. , Schroeder, Selle, Teran, & Stomakhin, 2015) to substantially reduce numerical error and artificial damping. This enables us to simulate substances with better accuracy than alternative methods in computer graphics.

Compared with liquids, the motion of dry sand is largely determined by the frictional contact between grains. In the theory of elastoplasticity, the modeling of large deformation (e.g., frictional contact) can be based on a constitutive law that follows the Mohr-Coulomb friction theory. Following Klár et al. (2016), we simulate dry sand based on the Saint Venant Kirchhoff (StVK) elasticity model combined with a Drucker-Prager non-associated flow rule. Plasticity models the material response as a constraint projection problem, where the feasible region (or yield surface) of the final material stress is restricted to be inside

$$tr(\boldsymbol{\sigma})c_F + \left\|\boldsymbol{\sigma} - \frac{tr(\boldsymbol{\sigma})}{d}\right\|_F \leq 0, \tag{8}$$

where d is the dimension and c_F is the coefficient of internal friction between grains of sand. The stress (i.e., the deformation gradient) of each sand particle is projected onto the yield surface to satisfy the second law of thermodynamics.

References

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science, 3*(5), 133-140.

Baillargeon, R. (2002). The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell handbook of childhood cognitive development, 1*, 46-83.

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science, 13*(3), 89-94. doi:10.1111/j.0963-7214.2004.00281.x

Batchelor, G. K. (2000). *An introduction to fluid dynamics.* Cambridge: Cambridge University Press.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2015). Humans predict liquid dynamics using probabilistic simulation. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, (pp. 172-177).

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences, 110*(45), 18327-18332.

Battaglia, P., & Pascanu, R. (2016). Interaction networks for learning about objects, relations and physics. *Advances in Neural Information Processing Systems*, (pp. 4502-4510).

Bridson, R. (2015). *Fluid simulation for computer graphics.* CRC Press.

Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. (2016). A compositional object-based approach to learning physical dynamics. *Proceedings of the 5th international conference on learning representations*, (pp. 1-15).

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

Clement, J. (1994). Use of physical intuition and imagistic simulation in expert problem solving. In D. Tirosh, *Human development, Vol. 6. Implicit and explicit knowledge: An educational approach* (pp. 204-244). Westport, CT: Ablex Publishing.

Cook, N. J., & Breedin, S. D. (1994). Constructing naive theories of motion on the fly. *Memory & Cognition, 22*(4), 474-493. doi:10.3758/BF03200871

De Kleer, J., & Brown, J. S. (1984). A qualitative physics based on confluences. *Artificial Intelligence, 24*(1-3), 7-83. doi:10.1016/0004-3702(84)90037-7

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences, 113*(34), E5072-E5081. doi:10.1073/pnas.1610344113

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. *Proceedings of the 37th annual conference of the cognitive science society*, (pp. 782-787).

Gilden, D. L., & Proffitt, D. R. (1994). Heuristic judgment of mass ratio in two-body collisions. *Perception and Psychophysics, 56*(6), 708-720. doi:10.3758/BF03208364

Grzeszczuk, R., Terzopoulos, D., & Hinton, G. (1998). Neuroanimator: Fast neural network emulation and control of physics-based models. *Proceedings of the 25th annual conference on computer graphics and interactive techniques*, (pp. 9-20).

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition, 157*, 61-76.

Hecht, H., & Proffitt, D. R. (1995). The price of expertise: Effects of experience on the water-level task. *Psychological Science, 6*(2), 90-95. doi:10.1111/j.1467-9280.1995.tb00312.x

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences, 8*(6), 280-285.

Hespos, S. J., Ferry, A. L., Anderson, E. M., Hollenbeck, E. N., & Rips, L. J. (2016). Five-month-old infants have general knowledge of how nonsolid substances behave and interact. *Psychological Science, 27*(2), 244-256. doi:10.1177/0956797615617897

Jiang, C., Schroeder, C., Selle, A., Teran, J., & Stomakhin, A. (2015). The affine particle-in-cell method. *ACM Transactions on Graphics (TOG), 34*(4), 51.

Jiang, C., Schroeder, C., Teran, J., Stomakhin, A., & Selle, A. (2016). The material point method for simulating continuum materials. *ACM SIGGRAPH 2016 Courses*, *24.* doi:10.1145/2897826.2927348

Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition, 14*(4), 308-312. doi:10.3758/BF03202508

Kaiser, M. K., Proffitt, D. R., Whelan, S. M., & Hecht, H. (1992). Influence of animation on dynamical judgments. *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 669-689. doi:10.1037/0096-1523.18.3.669

Kawabe, T., Maruya, K., Fleming, R. W., & Nishida, S. (2015). Seeing liquids from visual motion. *Vision Research, 109*, 125-138.

Klár, G., Gast, T., Pradhana, A., Fu, C., Schroeder, C., Jiang, C., & Teran, J. (2016). Drucker-prager elastoplasticity for sand animation. *ACM Transactions on Graphics (TOG), 35*(4), 103. doi:10.1145/2897824.2925906

Kotovsky, L., & Baillargeon, R. (1998). The development of calibration-based reasoning about collision events in young infants. *Cognition, 67*, 311-351.

Kozhevnikov, M., & Hegarty, M. (2001). Impetus beliefs as default heuristics: Dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin & Review, 8*(3), 439-453. doi:10.3758/BF03196179

Krist, H. (2000). Development of naive beliefs about moving objects: The straight-down belief in action. *Cognitive Development, 15*(3), 281-308. doi:10.1016/S0885-2014(00)00029-0

Krist, H., Fieberg, E. L., & Wilkening, F. (1993). Intuitive physics in action and judgment: The development of knowledge about projectile motion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(4), 952-966.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: current research and controversies. *Trends in Cognitive Sciences, 21*(10), 749-759.

Li, Y. W., Tedrake, R., Tenenbaum, J. B., & Torralba, A. (2018). Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566*.

Liang, W., Zhao, Y., Zhu, Y., & Zhu, S. (2015). Evaluating human cognition of containing relations with physical simulation. *Proceedings of the 37th annual conference of the Cognitive Science Society*, (pp. 1356-1361).

McAfee, E. A., & Proffitt, D. R. (1991). Understanding the surface orientation of liquids. *Cognitive Psychology, 23*(3), 483-514. doi:10.1016/0010-0285(91)90017-I

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research, 16*(3), 285-292.

Monaghan, J. J. (1992). Smoothed particle hydrodynamics. *Annual Review of Astronomy and Astrophysics, 30*, 543-574.

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fai-Fai, L. F., Tenenbaum, J., & Yaminds, D. L. (2018). Flexible neural representation for physics prediction. *Advances in Neural Information Processing Systems*, (pp. 8813-8824).

Museth, K., Lait, J., Johanson, J., Budsberg, J., Henderson, R., Alden, M., . . . Pearce, A. (2013). OpenVDB: an open-source data structure and toolkit for high-resolution volumes. *ACM SIGGRAPH 2013 courses* (p. 19). ACM.

Rebelsky, F. (1964). Adult perception of the horizontal. *Perceptual and Motor Skills, 19*(2), 371-374.

Runeson, S., Juslin, P., & Olsson, H. (2000). Visual perception of dynamic properties: Cue heuristics versus direct-perceptual competence. *Psychological Review, 107*(3), 525-555. doi:10.1037/0033-295X.107.3.525

Sanborn, A. N. (2014). Testing Bayesian and heuristic predictions of mass judgments of colliding objects. *Frontiers in Psychology, 5*(938).

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological review, 120*(2), 411-437.

Schwartz, D. L., & Black, J. B. (1996). Analog imagery in mental model reasoning: depictive models. *Cognitive Psychology, 30*(2), 154-219. doi:10.1006/cogp.1996.0006

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 116-136. doi:10.1037/0278-7393.25.1.116

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science, 5*(1), 185-199. doi:10.1111/tops.12009

Smith, K. A., Battaglia, P. W., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, (pp. 3426-3431).

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobsen, K. (1992). Origins of knowledge. *Psychological Review, 99*(4), 605-632.

Spelke, E. S., Katz, G., Purcell, S. E., Ehrlich, S. M., & Breinlinger, K. (1994). Early knowledge of object motion. *Cognition, 51*(2), 131-176. doi:10.1016/0010-0277(94)90013-2

Sulsky, D., Zhou, S., & Schreyer, H. (1995). Application of a particle-in-cell method to solid mechanics. *Computer physics communications, 87*(1-2), 236-352. doi:10.1016/0010-4655(94)00170-7

Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332*(6033), 1054-1059. doi:10.1126/science.1196404

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: can it facilitate? *International Journal of Human Computer Studies, 57*(4), 247-262.

Ye, T., Qi, S., Kubricht, J., Zhu, Y., Lu, H., & Zhu, S. C. (2017). The Martian: examining human physical judgments across virtual gravity fields. *IEEE Transactions on Visualization and Computer Graphics, 23*(4), 1399-1408. doi:10.1109/TVCG.2017.2657235

Zhu, Y., & Bridson, R. (2005). Animating sand as a fluid. *ACM Transactions on Graphics (TOG), 24*(3), 965-972. doi:10.1145/1073204.1073298

Zhu, Y., Jiang, C., Zhao, Y., Terzopoulos, D., & Zhu, S. (2016). Inferring forces and learning human utilities from videos. *IEEE conference on computer vision and pattern recognition (CVPR)*, (pp. 3823-3833).

Zhu, Y., Zhao, Y., & Zhu, S. (2015). Understanding tools: Task-oriented object modeling, learning, and recognition. *IEEE conference on computer vision and pattern recognition (CVPR)*, (pp. 2855-2864).