Physical Reasoning in Complex Scenes is Sensitive to Mass

by

Jessica B. Hamrick

Submitted to the Department of Electrical Engineering

and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

May 2012

Author ....
Department of Electrical Engineering and Computer Science
May 21, 2012

Certified by ....
Prof. Joshua B. Tenenbaum, Thesis Supervisor
May 21, 2012

Accepted by ....
Prof. Dennis M. Freeman, Masters of Engineering Thesis Committee

Physical Reasoning in Complex Scenes is Sensitive to Mass

by

Jessica B. Hamrick
Submitted to the
Department of Electrical Engineering and Computer Science

May 21, 2012

In Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Many human activities require precise judgments about the dynamics and physical properties – for example, mass – of multiple objects. Classic work suggests that people's intuitive models of physics in mass-sensitive situations are relatively poor and error-prone, based on highly simplified heuristics that apply only in special cases. These conclusions seem at odds with the breadth and sophistication of naive physical reasoning in real-world situations. Our work measures the boundaries of people's physical reasoning in mass-sensitive scenarios and tests the richness of intuitive physics knowledge in more complex scenes. We asked participants to make quantitative judgments about stability and other physical properties of virtual 3D towers composed of heavy and light blocks. We found their judgments correlated highly with a model observer that uses simulations based on realistic physical dynamics and sampling-based approximate probabilistic inference to efficiently and accurately estimate these properties. Several alternative heuristic accounts provide substantially worse fits. In a separate task, participants observed virtual 3D billiards-like movies and judged which balls were lighter. In contrast to the previous experiments, we found their judgments to be more consistent with simple, visual heuristics than a simulation-based model that updates its beliefs about mass in response to prediction errors. We conclude that rich internal physics models are likely to play a key role in guiding human common-sense reasoning in prediction-based tasks and emphasize the need for further investigation in inference-based tasks.

Thesis Supervisor: Joshua B. Tenenbaum
Title: Professor

Acknowledgments

First I would like to thank my thesis advisor, Josh Tenenbaum, for giving me a first chance and for believing in me before I did. Josh was the catalyst for this research, inspiring me to think about artificial intelligence and cognitive science in entirely new ways. His optimism, enthusiasm, and guidance always point me in the right direction.

I am indebted to Peter Battaglia, whom without this thesis would never have been possible. Whether he was patiently explaining the simplest of concepts or engaging me in joint coding sessions, he has been a friend and mentor every step of the way. His efforts to train me have profoundly shaped me as an aspiring cognitive scientist and our collaboration has had an enormous impact on this thesis.

I thank Max Siegel, Jonathan Scholz, Tomer Ullman, Tao Gao, Eyal Dechter, Philip Isola, John McCoy, and the rest of the "cogphysics" group for valuable insights and discussions. Many thanks also go to the rest of the MIT Computational Cognitive Science Group, especially Chris Baker and David Wingate, who accepted me as an undergraduate and who always give me useful feedback.

I am grateful to Gerry Sussman, my academic advisor, who showed me the elegance of symbolic programming and who was always interested to hear about my research. I came away from every meeting with him with a new point of view and many ideas. I also thank Laura Schulz, Leslie Kaelbling, and the other professors who have been role models and inspirations throughout the past five years.

My mom, Virginia Hamrick; my siblings, Ellie and David Hamrick; my sister-in-law, Rachel Hamrick; and the rest of the Chandler family, have been constant pillars of support before and during my time at MIT. Their love, guidance, and encouragement has never faltered.

Last, but far from least, I thank my dad, James Hamrick. Whether it was applying to MIT, enrolling in my first class in Brain and Cognitive Sciences, or "taking the high road", he always inspired my best decisions and was my loudest cheerleader. To his memory I dedicate this thesis.

This work was supported by Qualcomm, ONR MURI N00014-07-1-0937, ONR grant N00014-09-0124, ARO MURI W911NF-08-1-0242, and ONR MURI 1015GNA126.

Contents

List of Figures

Chapter 1

Introduction

In cognitive science, intuitive physics is defined as the ability to reason about and make judgments regarding the physical world: predicting the stability of structures, the path of a moving object, or the friction of a surface all fall under this domain. Humans frequently use their physical knowledge to better interact with the world around them, yet despite years of research in cognitive science (Caramazza, McCloskey, & Green, 1981; McCloskey, Caramazza, & Green, 1980; McCloskey, 1983; Todd & Warren, 1982; Gilden & Proffitt, 1989a, 1989b, 1994; Runeson, 1977; Sanborn, Mansinghka, & Griffiths, 2009; Fleming, Barnett-Cowan, & Bülthoff, 2010; Spelke, Breinlinger, Macomber, & Jacobson, 1992; Baillargeon, 1987, 2007, 1994), the underlying principles that guide these judgments and the acquisition of relevant knowledge is not well understood. Similarly in the world of artificial intelligence and robotics, machines are largely programmed for highly specific scenarios, lacking abstract knowledge of how the world actually works (Arimoto, 1999). While the representation of this knowledge for use in artificial intelligence has been extensively discussed (Hayes, 1978, 1984; Gardin & Meltzer, 1989; Forbus, 1981, 1983), there has been no general consensus on what form it should take.

Understanding the nature of physical knowledge is a crucial component in building intelligent machines that are able to interact with the world around them. Humans use this knowledge every day to avoid injury; for example, one would be wary of walking underneath a structure that seemed unstable for fear of it falling on them. This knowledge is used even in more mundane tasks, such as deciding where to place your cup of coffee such that it will be unlikely to fall off the table and spill. Machines will similarly benefit from such knowledge: take for example a demolition robot with a wrecking ball. The robot must be able to determine how much force it should use when swinging the ball; if it uses too much force, it could damage itself or surrounding buildings and people; if it uses too little force, it will not succeed at its task. Accurate physical knowledge may

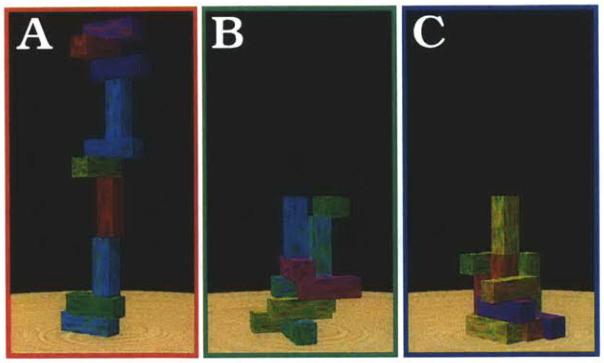


Figure 1.1: **Three towers of varying height and stability.** Hamrick et al. (2011) predicted the stability and fall direction of towers such as these using a probabilistic, simulation-based model of intuitive physics. A is clearly unstable, C clearly stable, while B (matched in height to C) is less clear.

also aid human and robot interaction: machines that can anticipate human physical abilities, limits, and goals will be more effective at assisting people with less risk of injury.

The kinds of judgments we consider here are those necessary to navigate, interact with, and constructively modify real-world physical environments. Consider the towers of blocks shown in Figure 1.1. How stable are these configurations, or how likely are they to fall? If they fall, in what direction will the blocks scatter? Where could a block be added or removed from the tower to significantly alter the configuration's stability?

1.1 Reasoning about mass

In addition to the geometric configuration of objects, physical parameters like mass, friction, elasticity, and shape can all affect how situations play out. Here we will place emphasis on one variable in particular: mass. While most previous literature on mass-based reasoning has debated whether humans use simple heuristics vs. visual invariants (Runeson, 1977; Jacobs, Michaels, & Runeson, 2000; Runeson, Juslin, & Olsson, 2000; Todd & Warren, 1982; Gilden & Proffitt, 1989a, 1989b, 1994; Hecht, 1996), Sanborn et al. (2009) demonstrated that human reasoning about mass could be explained by a Bayesian model of Newtonian collision dynamics. We believe this approach is

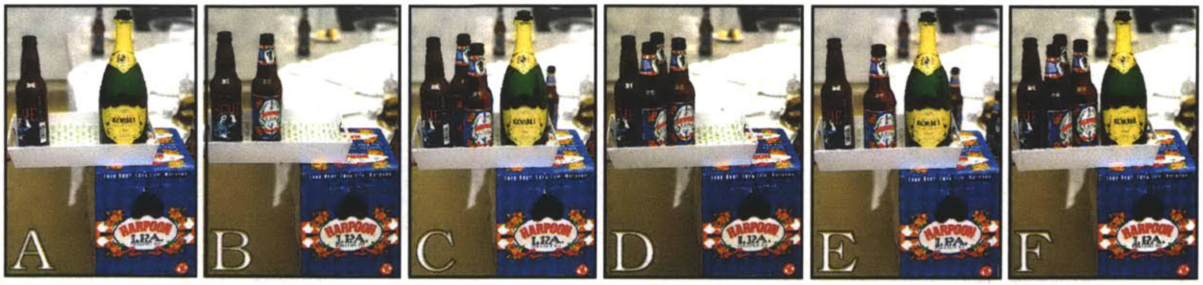


Figure 1.2: **Examples of stability and mass reasoning in real-world scenarios.** People frequently make judgments regarding mass and stability. For example, while at a party your friends might decide to balance various beverage containers. It is not difficult to determine which of these scenes are stable or plausible: A is probably stable, B is not stable, C is uncertain, D is definitely not stable, E might be stable, and F is probably not stable.

compatible with the ease and flexibility with which people reason about mass in every day life: given the complexity and diversity of the possible scenarios people encounter, a model based on realistic physical dynamics that can generalize to many situations seems appropriate.

Consider a party scenario in which your friends have decided to balance various beverage containers. Some possible (and impossible) versions of this scenario are depicted in Figure 1.2. If all the bottles are empty, which scenes are stable? Unstable? What if the champagne bottle is half-full? Completely full? People offer consistent judgments about these scenes and naturally adjust their judgments to account for changes in mass. For example, scene C may become more stable in the context of a full champagne bottle, but scene B will never be stable, regardless of the masses of the objects. Reasoning about such scenes may go in the opposite direction as well. It is clear that if scene F in Figure 1.2 is stable, the champagne bottle must be around four times heavier than a beer bottle. Similarly, if scene A is unstable, a natural inference is that the beer bottle is full and the champagne bottle is empty.

Reasoning about mass has applications beyond cognitive science. One of the classic goals of robotics and AI has been to create a robot that can assist humans by maneuvering through the world, picking up objects, and moving them to a new location. Incorporating into a robot an ability to reason about mass would be highly beneficial, enabling machines to form expectations of mass or behave rationally in mass-sensitive contexts. For example, a belief about an object's mass would enable a robot to choose a reasonable force with which to lift it. Similarly, if a robot can predict which structures are likely to be unstable due to mass, it can adjust its penalty function for bumping into things and more successfully navigate the environment.

1.2 Outline

Hamrick, Battaglia, and Tenenbaum (2011) hypothesized that humans have a rich, probabilistic, simulation-based model of intuitive physics that they use to make such judgments, and Sanborn et al. (2009) examined the ability of a Bayesian model of Newtonian collision dynamics to explain human judgments of mass. This thesis builds upon these works to examine both how people reason about complex, physical scenarios in the context of mass and how people can infer mass from the dynamics of such scenarios. Chapter 2 outlines previous work in the field of intuitive physics and presents evidence that supports humans' simulation-based intuitive physics. Chapter 3 develops a model of mass-sensitive predictions and through psychophysical observation, finds that it better explains people's behavior than competing, non-simulation-based models. Chapter 4 proposes a novel model of human physical property inference and compares its predictions to human judgments in a new psychophysical experiment. In brief, it cannot account for people's behavior, perhaps due to task difficulty or the availability of simple geometric heuristics. This highlights a rich area of future work on principled hybrids of shallow, heuristic reasoning strategies and rich simulation-based models. Finally, Chapter 5 concludes with a summary of this work's contributions and directions for future work.

Chapter 2

Theories of Intuitive Physics

Intuitive physics is a core domain of common-sense reasoning, developing early in infancy and remaining central in adult thought (Baillargeon, 2007). Yet, despite decades of research, there is no consensus on certain basic questions: What kinds of internal models of the physical world do human minds build? How rich and physically accurate are they? How is intuitive physical knowledge represented or used to guide physical judgments?

Classic research focused on the limits of human physical reasoning. One line of work argued that people's understanding of simple object trajectories moving under inertial dynamics was biased away from the true Newtonian dynamics, towards a more "Aristotelian" or "impetus" kinematic theory (Caramazza et al., 1981; McCloskey, 1983), yet no precise model of an intuitive impetus theory was developed. Studies of how people judge relative masses in two-body collisions concluded that humans are limited to making physical judgments based on simple heuristics, or become confused in tasks requiring attention to more than one dimension of a dynamic scene (Todd & Warren, 1982; Gilden & Proffitt, 1989a, 1989b, 1994). Neither the impetus accounts nor the simple one-dimensional heuristic accounts attempted to explain how people might reason about complex scenes such as Figure 1.1, or gave any basis to think people might reason about them with a high degree of accuracy.

2.1 An argument for a simulation-based account

Recent work in computer graphics and robotics demonstrates the power of realistic physical models: modern graphics demand extensive physical simulation to successfully fool visual perception into believing synthetic images are real, and roboticists have argued for the value of detailed physics models in interacting effectively with household objects (Toussaint, Plath, Lang,

& Jetchev, 2010). Given that people's capacity for physically sensible perception and action far exceeds the best machine graphics or robotics, it is plausible that realistic physical models may be embedded in human perceptual and cognitive faculties.

We hypothesize that humans can make physical judgments using an internal generative model that approximates the principles of Newtonian mechanics applied to three-dimensional solid bodies (Figure 2.1). They use this model to forward-simulate future outcomes given beliefs about the world state, and make judgments based on the outcomes of these simulations. We believe that only by positing such rich internal models can we explain how people are able to perform complex everyday tasks like constructing and predicting properties of stacks of objects, balancing or stabilizing precariously arranged objects, or intercepting or avoiding multiple moving, interacting objects. Tenenbaum, Griffiths, and Niyogi (2007) propose the use of generative causal grammars in modeling intuitive theories, emphasizing the "infinite use of finite means" that a generative grammar entails. The number of possible physical scenarios that people can encounter is infinite; we therefore also seek an approach that can handle an infinite number situations with a finite set of rules.

The notion of forward-simulation is not completely novel: it builds on evidence supporting the importance of rich forward models for vision and action. Yuille and Kersten (2006) suggest that visual perception applies forward generative models of high-level scene representations to disambiguate and efficiently code low-level information, consistent with "predictive coding" theories of perceptual cortical processing (Rao & Ballard, 1999; Lee & Mumford, 2003). Wolpert and Kawato (1998) and others highlight the heavy involvement of forward sensory models of action consequences for open-loop motor control. Other research reports that rats plan navigational behaviors by rolling out potential future outcomes of possible path choices (Johnson & Redish, 2007).

In the physical situations human regularly encounter, the core principles are few and simple: knowledge about gravity, object solidity, and momentum may often be sufficient to determine outcomes from a given initial condition. These core concepts develop early and systematically in

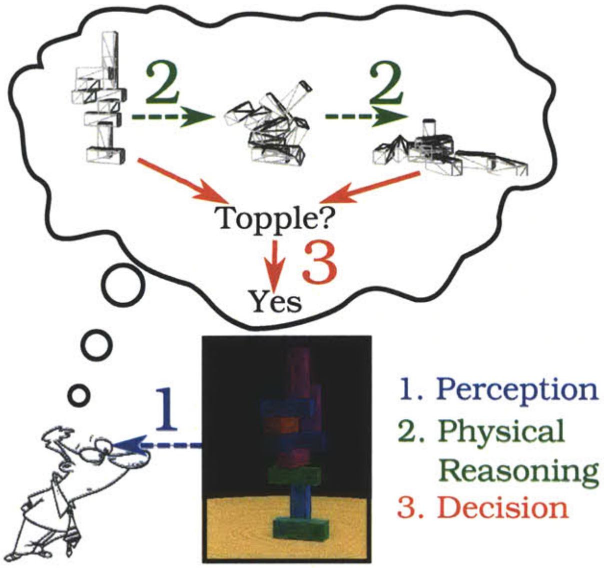


Figure 2.1: **Simulation-based theory of intuitive physics.** Our theory of how people reason about the physical world, as proposed in Hamrick et al. (2011). People first *perceive* the world and form an internal belief about what objects are there and where they are. They then simulate the future through a process of *physical reasoning* approximating Newtonian physics. Finally, they make a *decision* about what will happen (such as whether the tower of blocks will fall down) by comparing the observed scene to the predicted scene.

infants. Spelke et al. (1992) showed that 4-month old infants understand object solidity and continuity. Baillargeon (1987, 2007) reports that one-year old children reason about the stability of simple and irregularly shaped objects, and further argues that infants transition from using qualitative, all-or-none heuristics when reasoning about their physical world, to using a more quantitative, continuous, and general representation (Baillargeon, 1994). The early development and use of these foundational physical concepts supports our hypothesis that accurate core physics knowledge is leveraged to support complex judgments.

2.2 Evidence of generative physics knowledge

Hamrick et al. (2011) proposed a simulation model of "intuitive mechanics" that made Newtonian-based physical predictions in the context of perceptual uncertainty. The physical laws of the internal models we proposed are essentially deterministic, but people's judgments are probabilistic.

Capturing that probabilistic structure is crucial for predicting human judgments precisely and explaining how intuitive physical reasoning successfully guides adaptive behavior, decision-making and planning in the world. We can incorporate uncertainty in several ways. Objects' positions and velocities and their key physical properties (e.g., mass, coefficients of friction) may only be inferred with limited precision from perceptual input. People may also be uncertain about the underlying physical dynamics, or may consider the action of unobserved or unknown exogenous forces on the objects in the scene (e.g., a gust of wind, or someone bumping into the table). We represented these sources of uncertainty in terms of probability distributions over the values of state variables, parameters or latent forces in the deterministic physical model. By representing these distributions approximately in terms of small sets of samples, uncertainty can be propagated through the model's physical dynamics using only analog mental simulations. Thus a resource-bounded observer can make appropriate predictions of future outcomes without complex probabilistic calculations. Even though these simulations may approximate reality only roughly, and with large numbers of objects may only be sustainable for brief durations, they can still be sufficient to make useful judgments about complex scenes on short time-scales.

Through several experiments, we demonstrated the effectiveness of this model in explaining peoples judgments on two physical tasks: the stability of towers of blocks (such as those in Figures 1.1 and 2.4A) and the direction such towers would fall.

Stability paradigm We asked participants to judge the stability of towers of blocks by asking them, "Will this tower fall?". We ran several variations of this experiment: *feedback* (in which a movie of the tower falling or not was presented on each trial after participants entered their response), *no feedback* (in which participants only received feedback during a training period), and *same height* (in which all stimuli were of the same height). In each of these conditions, we found high correlations between the simulation-based model and human judgments. In particular, the model demonstrated robustness in the *same height* condition, in which the best non-simulation cue (tower height) was controlled for.

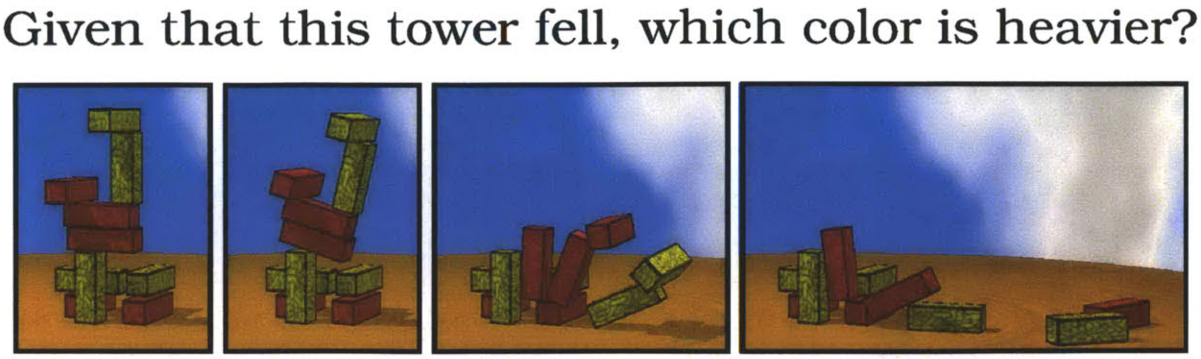


Figure 2.2: **Mass inference task for an unstable tower.** All of the red blocks have the same mass m_{red} and all of the yellow blocks have the same mass m_{yellow}. Given these frames of the tower falling, it is relatively straightforward to infer that m_{yellow} is greater than m_{red}.

Direction paradigm We presented participants with all unstable towers and asked them to judge the direction that the tower would fall in. The model was again highly predictive of peoples judgments, especially when the highest-variance towers were excluded from analysis.[1]

These two paradigms offer distinctly different tasks, both of which people have no difficulty performing. Importantly, the simulation-based model is able to explain people's judgments in both the stability and direction experiments with minimal modification. In contrast, heuristic-based explanations do not generalize well, requiring a new set of cues for each task. These results, in combination with several other recent lines of research (Zago & Lacquaniti, 2005; Fleming et al., 2010; Sanborn et al., 2009), provide strong evidence for the hypothesis that approximate Newtonian principles and simulation underlie human judgments about dynamics and stability.

2.3 Do people take mass into account?

Previous work that has investigated human reasoning about mass has focused on the relative roles of image invariants versus biases and heuristics in simple mass inference tasks. The invariants theory, or "kinematic specification of dynamics" (KSD), argues that people directly perceive object

[1] In circular statistics, the mean of a distribution conveys increasingly less information as the variance of the distribution increases. Therefore, the fall direction for high-variance towers is ill-defined and it is unreasonable to expect consistent predictions on these towers from people and the model.

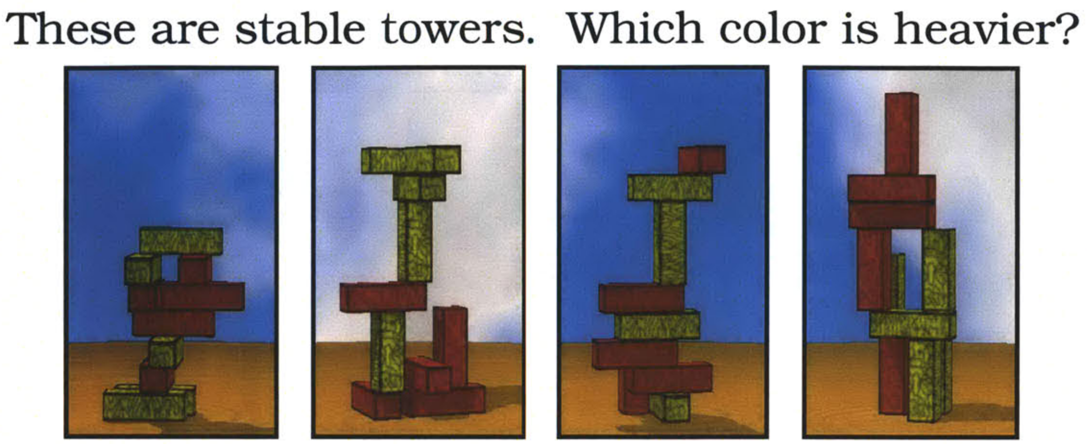


Figure 2.3: **Mass inference task for stable towers.** As in Figure 2.2, all of the red blocks have the same mass m_{red} and all of the yellow blocks have the same mass m_{yellow}. One of the colors *must* be heavier for these towers to be stable. For each tower, which color is heavier? People typically come up with the correct answers: yellow, red, yellow, and yellow.

dynamics – including mass – because there is a one-to-one correspondence between the objects' kinematic motion and the underlying properties (Runeson, 1977; Runeson et al., 2000; Jacobs et al., 2000). In contrast, the heuristics theory argues that people use simple perceptual cues, such as object velocity or amount of ricochet following a collision, to guide their judgments of mass (Todd & Warren, 1982; Gilden & Proffitt, 1989a, 1989b, 1994). Despite a large body of work arguing for both accounts, there has been little consensus on which strategy people actually use.[2]

These approaches to studying human inferences of mass have focused on specialized cases of of two point masses colliding in only one or dimensions. However, in reality, people rarely reason about low dimensional ideal collisions. Real object surfaces suffer from imperfections, preventing perfect collisions; similarly, supporting surfaces and platforms are not perfectly smooth, injecting a large degree of noise. This chaos is perhaps part of the appeal of games like *bocce*, in which players attempt to throw balls as close to another ball as possible. As the game is frequently played on grass or soil, the balls bounce and roll in highly unpredictable ways.

Sanborn et al. (2009) frames intuitive physics as a kind of probabilistic Newtonian mechanics in which uncertainty about latent variables gives rise to uncertain predictions of physical outcomes.

[2] See Hecht (1996) for further commentary on this debate

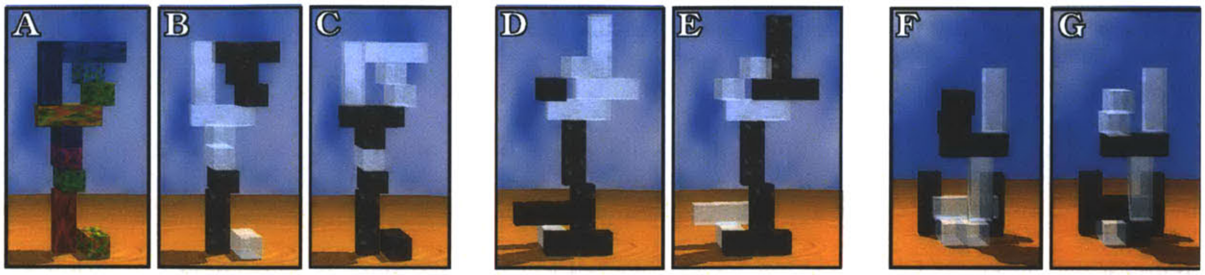


Figure 2.4: **Mass-sensitive towers of varying stability and fall direction.** Hamrick et al. (2011) showed participants towers such as A and asked them to predict whether the towers would fall or in what direction they would fall. This work modifies stimuli like A to have two types of blocks: black stone (very heavy) and white, translucent plastic (very light). Towers B and C have the same *geometric configuration* as A but different *mass configurations*; B's mass configuration makes it very unstable, while A is very stable. Towers D and E have the same geometric configuration, but D will fall to the left and E will fall to the right. Towers F and G also have the same geometric configurations, but F is unstable while G is stable.

In contrast to the previous work on mass inference, they showed that such a framework is able to explain human judgments of mass in two-body collisions. These results, combined with the initial indications from Hamrick et al. (2011), point to the possibility that people may not use simple heuristics or invariants to reason about mass, but instead use a much richer, flexible model approximating Newtonian dynamics to reason about mass.

We emphasize that people reason about mass in contexts other than collisions. Perhaps the most canonical example is the aptly-named paperweight: a heavy object rests on paper, pinning it down so that will not be blown away by air currents. Consider more complex scenarios, such as the unstable tower in Figure 2.2 and the stable towers in Figure 2.3. Inferring which color block is heavier is much more difficult than inferring the purpose of a paperweight and requires significant logical reasoning, but it is not impossible.

Because objects' masses play a key role in real physical dynamics, we further hypothesize that people's predictive judgments factor in mass in a physically appropriate manner. For a particular configuration of block positions and orientations (henceforth "geometric configuration"), we can choose different blocks to be heavy or light. We refer to this choice of block type as the "mass configuration". In particular, there are cases in which different mass configurations of the same geometric configuration induce different physical properties. For example, towers B & C and F & G in Figure 2.4 have the same geometric configurations, yet the mass configurations of B and F are very unstable, while the mass configurations of the C and G are clearly balanced. Similarly, towers D & E also have the same geometric configuration, but the different mass configurations cause tower D to fall towards the left and tower E to fall towards the right. If such identical geometric configurations have different physical outcomes and elicit different human judgments as well, then it is clear that people are incorporating mass into their reasoning process.

The main innovation of this work is twofold: (1) to capture physical knowledge with a three-dimensional and realistic object-based physics simulation and to implement probabilistic inference using sample-based approximations; and (2) to examine physical reasoning in both *mass-sensitive prediction* and *mass inference* tasks in the context of *information from structure* and *information from collision dynamics*. Our more general framing allows us to test whether and how a probabilistic-Newtonian framework can scale up to explain intuitive physical reasoning in mass-sensitive scenes such as Figures 2.4, 2.2, and 2.3.

Chapter 3

Mass-Sensitive Predictions

3.1 Simulation-based prediction model

The model defined here is identical to the model from Hamrick et al. (2011). However, as it is a crucial foundation for the results of the experiments in Section 3.3 and the inference model in Chapter 4, we find it relevant to explain it here as well.

3.1.1 Definition

We frame human physical judgments using a probabilistic model observer (Figure 3.1) that combines three components: *perception*, *physical reasoning*, and *decision*. The perception component defines a mapping from input images to internal beliefs about the states of objects in a scene. The physical reasoning component describes how internal physics knowledge is used to predict future object states. The decision component describes how these predicted states are used to produce a desired property judgment. Uncertainty may enter into any or all of these components. [1]

Perception Our specific experimental focus is on judgments about dynamic events with towers of blocks (Figure 2.4), so the relevant object states S_t are the locations and orientations of all blocks in the tower at time t.

Perceptual uncertainty People's inferences about the geometry have uncertainty due to a variety of factors, including visual occlusion, perceptual noise, and working memory imprecision. We model their perceptual beliefs about the scene as posterior inference, and quantify the scale

[1] For simplicity, in this work we have modeled uncertainty only in the perception component, assuming that observers compute a noisy representation of objects' positions in the three-dimensional scene. Similar noise distributions applied to objects' states could also represent other sources of uncertainty, such as unknown latent forces in the world that might perturb the objects' state or uncertainty about specific physical dynamics. Here we do not distinguish these sources of uncertainty but leave this as a question for future work.

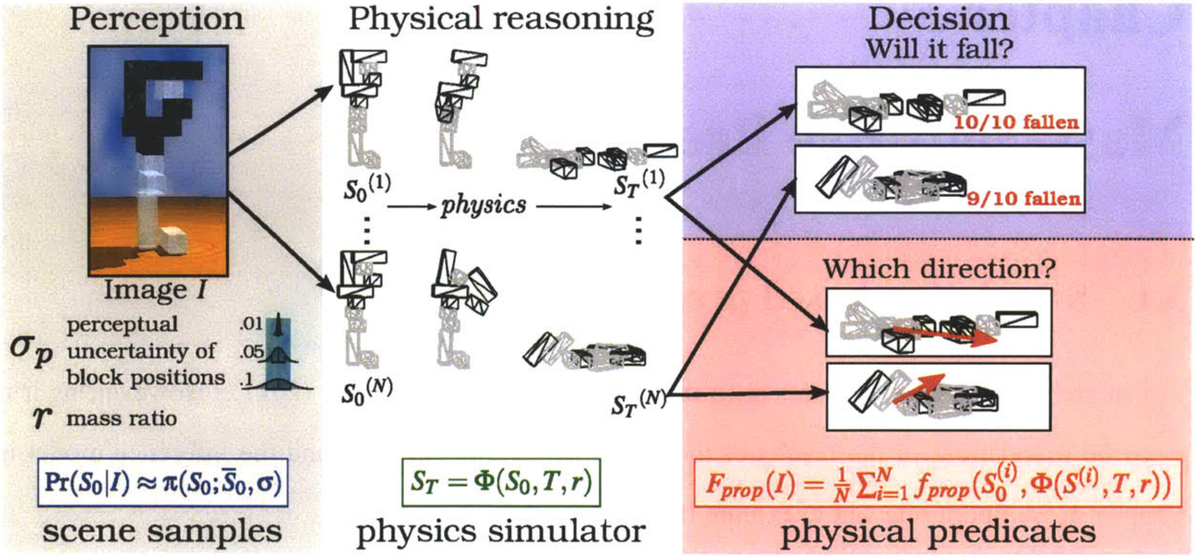


Figure 3.1: **Prediction model schematic.** Our model has 3 components, perception, physical reasoning, and decision. During perception, an uncertain belief about the tower is inferred from an image. During physical reasoning, tower samples are drawn from this belief distribution, and a physical simulation is applied to each. To make decisions about physical properties, the simulation outcomes are evaluated and averaged.

of their horizontal positional uncertainty in their perception of each object's location using the parameter σ_p. When $\sigma_p = 0$, the model's outputs are deterministic and correspond to physical ground-truth judgments.

Mass ratio The relationship between heavy and light objects (black and white blocks, respectively, in Figure 2.4) is represented by r, the heavy-to-light mass ratio. When $r = 1$, the model treats all objects as if they have the same mass (as in Hamrick et al. (2011)).

Physical reasoning The effect of Newtonian physics over time on the tower, which includes gravitational forces, elastic collisions, and transfer of energy, is represented by the function $\Phi(\cdot)$, which inputs S_t and temporal duration T, and outputs the subsequent state $S_{T+t} = \Phi(S_t, T, r)$.

Estimating a physical tower property means predicting the future based on the present. In principle, deterministic physics implies that knowledge of $\overline{S}_0$ and $\Phi(\cdot)$ is sufficient to predict future physical properties perfectly. However, a realistic observer does not have direct access to tower states, $\overline{S}_0$, so must rely on uncertain perceptual inferences to draw beliefs about the tower. The observer forms beliefs about $\overline{S}_0$ conditioned on an image, I, and represents these beliefs, S_0, with the distribution, $\Pr(S_0|I)$. Applying physics to the inferred initial state S_0 induces a future state $S_T = \Phi(S_0, T, r)$ with distribution $\Pr(S_T|I)$.

Decision There are many possible judgments that could be made regarding the change between the current observed tower state and the estimated future tower state. While we only examine two kinds of judgments, we emphasize that our model can generalize to any type of judgment that involves comparing initial and future scene states. The judgments involve reasoning about the future state S_T of a tower first observed at $t = 0$:

1. What proportion of the tower will fall, $f_{fall}(I)$?
2. In what direction will the tower fall, $f_{dir}(I)$?

Predicting a physical property means computing: $f_{prop}(S_0, S_T) = f_{prop}(S_0, \Phi(S_0, T, r))$. To make decisions about physical properties, the observer computes the expectation:

$$\overline{F}_{prop}(I) = \mathrm{E}[f_{prop}(S_0, \Phi(S_0, T, r))]_I = \int_{S_0} f_{prop}(S_0, \Phi(S_0, T, r)) \cdot \Pr(S_0|I)\ \mathrm{d}S_0 \qquad (3.1)$$

which represents the model observer's estimate of the physical property f given I, integrating out the perceptual uncertainty in the initial state S_0.

3.1.2 Algorithm

As illustrated in Figure 3.1, we approximate Equation 3.1 though a Monte Carlo simulation procedure that (for a given mass ratio r) draws N "perceptual" samples ${S_0}^{(1,\ldots,N)} \sim \pi(S_0; \overline{S}_0, \sigma_p)$, simulates physics forward on each sample to time T, and computes the mean value for physical property f across their final states:

$$\overline{F}_{prop}(I) \approx F_{prop}(I) = \frac{1}{N} \sum_{i=0}^{N} f_{prop}({S_0}^{(i)}, \Phi({S_0}^{(i)}, T, r)). \qquad (3.2)$$

Scene samples We model $\Pr(S_0|I)$ in a way that reflects perceptual uncertainty and the principle that objects cannot interpenetrate, without committing to particular assumptions about perceptual inference or representations. We approximate $\Pr(S_0|I) \approx \pi(S_0; \overline{S}_0, \sigma_p)$, where $\pi(\cdot)$ is a composition of two terms: a set of independent Gaussian distributions for each block's x and y positions, with variance ${\sigma_p}^2$ and mean centered on the corresponding block positions in the true world state $\overline{S}_0$, followed by a deterministic transform that prevents blocks from interpenetrating. This is clearly a simplified approximation to the observer's perceptual distribution, but it serves as a reasonable starting point, and a place where more sophisticated vision models can be interfaced with our approach in future work.

Physics simulator The simulations[2] depend on 4 parameters: movement threshold, $M = 0.1$m, the distance a block must move before it is considered to have "fallen"; timescale of the simulation, $T = 2000$ms, defined above; σ_p, the perceptual uncertainty; and r, the mass ratio of heavy to light blocks. We found $\sigma_p = 0.04$ and $r = 4$ to provide reasonable fits to all conditions but we explore the effects of varying σ_p and r below (Figure 3.4).

Pilot analyses determined that our results were insensitive to changes nearby the chosen M and T values, while both σ_p and r had substantial effects in comparison with peoples' judgments.

Physical predicates We quantify degree of stability, f_{fall}, as the proportion of a tower that remains standing following the application of physics for duration T. This definition matches the objective notion that a tower that entirely collapses should be judged less stable than one for which a single block teeters off. The fall direction, f_{dir}, is measured as the angle of the mean position of those *heavy* blocks that have fallen.

[2] Our implementation of the physical simulation used the Open Dynamics Engine (Smith, 2009), a standard computer physics engine, which, critically, allows precise simulation of rigid-body dynamics with momentous collisions.

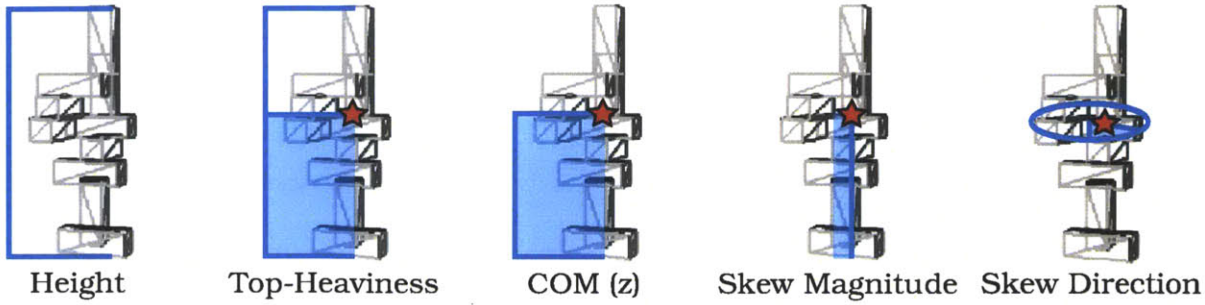


Figure 3.2: **Geometric prediction heuristics**. *Height* is the difference between the base of the lowest block and the top of the highest block along the vertical (z) dimension. *Top-heaviness* is the ratio of the $z-$coordinate of the center of mass to the height of the tower. *COM (z)* is the $z-$coordinate of the center of mass. *Skew magnitude* and *skew direction* are (respectively) the radius and angle in cylindrical coordinates of the tower's center of mass.

3.2 Visual and geometric heuristics for prediction

To evaluate alternative explanations of humans' judgments, we test whether several mass-sensitive heuristics (Figure 3.2) may account for their responses. Here, the tower's center of mass is a weighted average of the block positions, where the weights are proportional to the block masses. We define the tower's center of mass in cylindrical coordinates (ρ, θ, z), where the z-axis is vertical, and examine the following heuristics:

- H_h (height): the geometric height of the tower. This heuristic is not mass-sensitive but was included in our analyses because it was the best-performing stability heuristic for uniform mass towers (Hamrick et al., 2011).

- H_z (center of mass, $z-$coordinate): the vertical distance of the center-of-mass from the ground.

- H_{th} (top-heaviness): the ratio of the H_z to H_h; essentially, the vertical distribution of mass.

- H_ρ (skew magnitude): how far out in the $x-y$ plane the center of mass is from the tower's geometric center.

- H_θ (skew direction): the angle of the center of mass with respect to the tower's geometric center.

The H_h, H_ρ, and H_z heuristic measures are inversely proportional to physical stability, e.g. tall towers tend to be less stable. For clarity, we negated the values of H_h and H_ρ to instead reflect a proportional relationship; this does not affect the correlations beyond changing their sign. We used the inverse of the center of mass heuristic, $\frac{1}{H_z}$, which provided a better correlation than just H_z itself.

3.3 Mass-sensitive prediction experiments

To examine the effect of mass in human judgments of physical scenarios, we modified the "stability" and "direction" experiments from Hamrick et al. (2011) to use mass-sensitive stimuli. Instead of towers of identical blocks, we presented participants with towers of heavy and light blocks.

3.3.1 Methods

Participants

20 adult participants were recruited from the MIT BCS human subject pool, and each gave informed consent in accordance with MIT's IRB. They were compensated \$10/hour, and performed 1 hour each.

Apparatus

Participants viewed an LCD monitor (refresh 60Hz) attached to a standard desktop PC, from an approximate distance of 0.6 meters. All stimuli were rendered in 3D using Panda3D (CMU Entertainment Technology Center, 2012) and physics simulations were computed at 1000Hz using the ODE physics engine (Smith, 2009). Animations of the physical dynamics were played back at

2x speed. Participants submitted response judgments by depressing keyboard keys (as in Section 3.3.3) or by adjusting an indicator with the mouse (as in Section 3.3.4).

All trials depicted a 3D scene that contained a circular 3m radius "ground disk", and a tower of 10 blocks (each block 20 × 20 × 60 cm) placed at the disk's center. The ground was textured with a wood grain pattern (e.g., Figure 3.3).

Trial structure

All trials had 3 phases: *stimulus*, *response*, and *feedback*.

Stimulus phase. The *stimulus* phase was 3000ms, during which participants passively viewed the tower of blocks from a camera that rotated around the tower to $+90°$, then to $-90°$, and back to the initial viewing angle. The camera radius was 10m, and the field of view was 40°. No physics simulations were applied, so the only image motion was due to camera rotation. After 3000ms, a cylindrical "occluder" descended vertically over 50ms and rested on the ground plane to obscure the tower from view; this ended the stimulus phase.

Response phase. The *response* phase then began immediately, and was not limited in time, but ended once the participant depressed a response key (stability experiment) or clicked the left mouse button (direction experiment).

Feedback phase. The *feedback* phase began immediately after the response phase and lasted 1500ms. During feedback, the occluder ascended vertically out of sight over 50ms, and for the remaining 1450ms the tower was visible. During feedback, physics was turned "on", which meant gravitational acceleration of -9.8m/s was applied to the tower's blocks. Physics caused some towers to collapse, while some remained standing – this indication gave participants feedback about the accuracy of their response. Additionally, in the "stability experiment" (Section 3.3.3) the ground pattern was shaded green if the tower fell, and blue if the tower remained standing. After 1500ms, the tower was removed from the scene and the next trial's stimulus phase began.

Before the actual experiment, participants performed 20-trial "training" session to be familiarized with the task. This training session was identical to the actual experiment, with the exception that separate stimuli were used. All experimental conditions were composed of 384 trials of 4 subsessions (96 trials per subsession). In each subsession, 48 different geometric configurations were displayed twice with two different mass configurations; these were randomized in order and repeated across subsessions such that there were at least 5 trials in between identical geometric configurations. Participants were not told that towers were repeated, and the viewing angle of the camera was randomized across the range of 0° to 360° for each trial.

Tower stimuli

Each tower was constructed by a stochastic process in which 10 blocks were sequentially given positions and orientations that resulted in a stable or unstable stack of blocks. Specifically, each block's placement satisfied two constraints: 1) its center must reside within the 60×60 cm length and width of the tower, and 2) it must be "locally stable", meaning that the block is supported by the blocks beneath it (however adding more blocks on top could cause it to fall). We then scored each tower's "true stability" by simulating physics (i.e. gravity) and measuring whether any blocks in the tower fell within 2000ms – those that had blocks fall were deemed unstable, and those from which no blocks fell deemed stable. Stimuli from the "stability" experiment can be seen in Figure 3.3 and stimuli from the "direction" experiment in Figure 3.5.

In all stimuli, half of the blocks were heavy (rendered with a black stone material) and half were light (rendered with a white, translucent plastic material). In the physical simulation, the blocks were assigned a mass ratio of 10 (stone) to 1 (plastic); the actual densities were $1700\frac{kg}{m^3}$ and $170\frac{kg}{m^3}$, respectively. Participants were told that the stone blocks were much heavier than the plastic block, but were not told the explicit mass ratio.

3.3.2 Analysis

On each trial, we present a tower with state $\overline{S}_0$. The human observer responds with the stability property R_{fall} or R_{dir}, depending on condition. We computed the mean R across participants for each tower as their collective physical property judgment (Figs. 3.4, 3.6). Similarly, we computed the mean across model samples, F_{fall} or F_{dir}, for each tower as its physical property judgment. We performed correlation analyses using Pearson's correlations, circular correlations, and partial correlations, as noted. All correlations were computed using a bootstrap analysis (repeated 1000 times, sampling with replacement) over individual human and model judgments. We computed correlations between mean human and model judgments for each bootstrap sample, and from these computed the mean correlations as well as standard error.

We performed an additional bootstrap analysis in the stability experiment to judge the effects of mass on participant responses. In this analysis, we resampled individual human and model responses with replacement, found the mean R_{fall} and F_{fall} for each mass configuration, and took the difference between means for mass configurations of the same geometric configuration, g_i. This process was repeated 1000 times, generating an approximately Gaussian judgment-difference distribution, from which we computed the probability that the two mass configurations were judged differently: $p_{diff}(g_i) = 1 - \Pr(|R_{diff}(g_i)| \leq 0)$ (or $1 - \Pr(|F_{diff}(g_i)| \leq 0)$, in the case of the model). For a pair of mass configurations to be considered significantly different, we required that $p_{diff} > 0.95$.

3.3.3 Experiment 1: "Stability" prediction

Methods

The first experiment asked participants to judge whether towers were stable or unstable. The 96 tower stimuli were randomly selected such that 50 were stable and 46 were unstable. The two mass configurations per geometric configuration ("pairs") were chosen such that there were 18 "stable pairs" (where both mass configurations resulted in stable structures), 16 "unstable pairs" (where both mass configurations resulted in unstable structures), and 14 "mixed pairs" (where one mass configuration induced a stable structure and the other an unstable structure).

On each trial participants made graded responses to the question, "Will this tower fall?", by

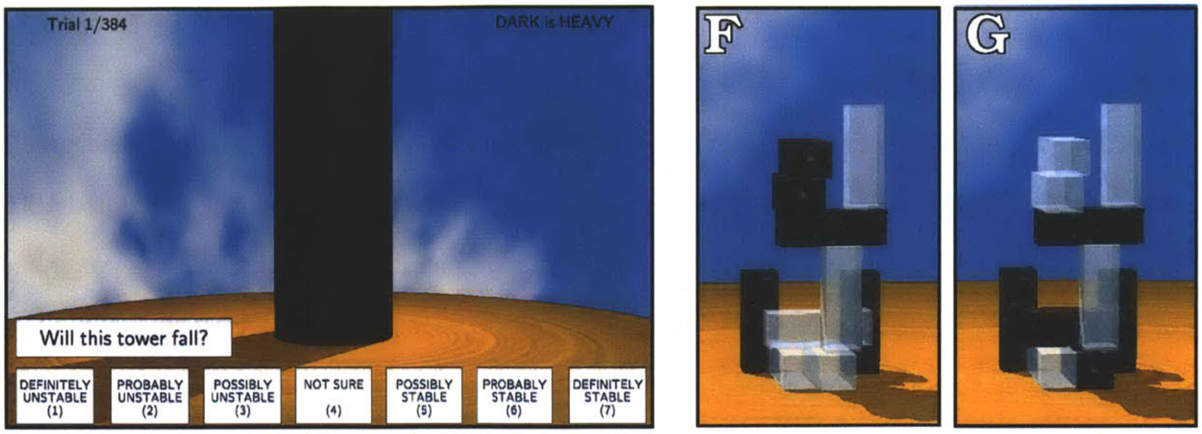


Figure 3.3: **Screenshot of the mass-sensitive "stability" prediction experiment.** Participants looked at stimuli such as Towers F and G, where the masses of the blocks (dark, heavy stone vs. light, translucent plastic) affected the stability of the structures. They answered the question, "Will this tower fall?", on a 1-7 scale ranging from "definitely unstable" (1) to "definitely stable" (2), as shown in the screenshot on the left.

pressing keys on a 1-7 scale to indicate degrees of confidence between "definitely unstable" (1) to "definitely stable (7) (Figure 3.3).

Additionally, participants were shown 10 "example" trials prior to the training period. These towers were specifically constructed to convey information about the 10:1 mass ratio without explicitly stating it.

Results

To determine whether observers use internal physics knowledge when making stability judgments, we computed the correlation between participants' judgments, R_{fall} ($n = 10$), our models' predictions, F_{fall}, and the geometric heuristics (Figure 3.4).

Upper bound To set an upper bound on how well the model can do, we performed a bootstrap analysis on human data, correlating the first half of the data with the second half on each bootstrap repetition. We found a correlation coefficient of 0.96 ± 0.01, similar to the human v. human correlation found in Hamrick et al. (2011).

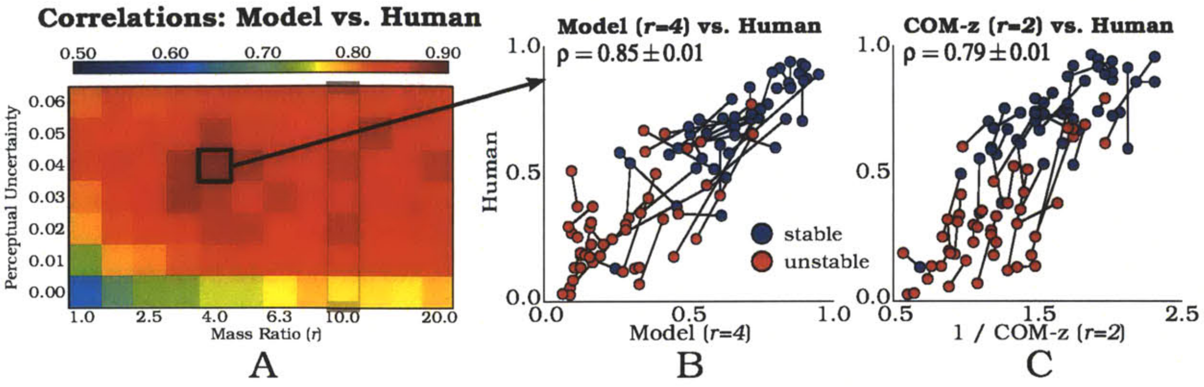


Figure 3.4: **Results of the "stability" prediction experiment.** A shows correlations between the simulation-based model and people for different values of r (mass ratio) and σ_p (perceptual uncertainty). B and C show scatter plots of the best simulation-based model and best heuristic vs. people. In these plots, the two mass configurations of the same geometric configuration are connected by a line. The red dots indicate unstable mass configurations and the blue dots indicate stable mass configurations.

Overall stability judgments In comparison with the ground truth model, which corresponded to zero uncertainty ($\sigma_p = 0$, $r = 10$), the correlation coefficient was 0.75 ± 0.01 (standard error, SE). The value of σ_p for which the model best-predicted human responses was 0.04; we note that $\sigma_p = 0.04$ was also the best uncertainty value in the uniform-mass stability and direction experiments from Hamrick et al. (2011). The mass ratio for the model that best-predicted human responses was 4:1 and yielded a correlation of 0.85 ± 0.01; however, when the model used the true mass ratio, 10:1, it did not yield a significantly different correlation ($\rho = 0.84 \pm 0.01$). Figure 3.4A shows the results of varying the perceptual uncertainty and mass ratio, and Figure 3.4B shows the performance of the $r = 4$ model.

Mass-sensitivity In the uniform-mass case ($r = 1$), the model does not predict peoples judgments as well, with a correlation of 0.81 ± 0.01. Although there is not a large difference in correlation between the $r = 4$ and $r = 1$ models, a partial correlation analysis suggests that the $r = 4$ model may be more predictive of peoples judgments. Removing the effect of the $r = 1$ model, we find a model/human correlation of 0.48 ± 0.05. If we remove the effect of the $r = 4$ model, we find a uniform-mass model/human correlation of 0.18 ± 0.07. These partial correlations lend support to the hypothesis that the mass-sensitive model is able to explain human judgments where the uniform-mass model cannot. Furthermore, the model ($r = 4$) was able to predict the magnitude of differences in people's judgments moderately well, with a correlation of 0.37 ± 0.08. The uniform-mass model could not predict these differences at all ($\rho = 0.13 \pm 0.11$).

To further quantify this effect of mass on human judgments, we performed a bootstrap analysis of human and model judgment differences to determine which pairs of mass configurations people judged differently (see 3.3.2). Humans judged 23/48 pairs differently and the model ($r = 4$, $\sigma_p = 0.04$) judged 21/48 pairs differently. The model and humans agreed on 15 of these pairs and judged the sign of the differences the same way on all but two. Overall, the model and humans agreed that 32/48 pairs were either significantly different or not significantly different. The uniform-mass model, in contrast, judged 5/48 pairs differently. The uniform-mass model *shouldn't* judge any of the pairs differently as it is not sensitive to mass; we suspect that this discrepancy is due to the highly nonlinear effects of the perceptual noise, σ_p, or the small number of model samples ($N = 20$). Additionally, although the uniform-mass model agreed with people on one significantly different pair, they disagreed about the sign of the difference. Thus, it is clear that people do take mass into account when judging tower stability. Furthermore, a mass-sensitive model is much more consistent with differences in human judgments than a mass-insentivie model.

Heuristics To test the possibility that people use simple heuristics rather than a richer model, we compared stability judgments predicted by the heuristics introduced in Section 3.2 with human judgments. For $r = 2$, which yielded the best heuristic correlations, the correlation coefficient between humans and: 1) tower height, H_h, was 0.64 ± 0.01, 2) center of mass (z-coordinate), H_z, was 0.79 ± 0.01, 3) top-heaviness, H_{th}, was 0.59 ± 0.01, and 3) tower skew magnitude, H_ρ, was 0.19 ± 0.01. Only the correlation for H_z was comparable to the model correlations, the others likely reflect inherent dependencies between the physical properties the heuristics represent and the actual physical stability of the towers. Figure 3.4C shows the relationship between people's judgments and center of mass.

Interestingly, the height heuristic, which is *not* sensitive to mass (but was the best heuristic found in Hamrick et al. (2011)), performs moderately well but is still much worse than either the model or the best heuristic. It is possible that because height is such a strong predictor in the uniform-mass case, people still rely on it to some extent.

To decouple the model predictions from the heuristics' effects, we computed the partial correlations between human data and the heuristics, controlling out the model's predictions. The partial correlations between humans and: 1) tower height, H_h, was 0.31 ± 0.04, 2) center of mass ($z-$coordinate), H_z, was 0.38 ± 0.04, 3) top-heaviness, H_{th}, was 0.16 ± 0.04, and 3) tower skew magnitude, H_p, was 0.06 ± 0.04. These correlations suggest that top-heaviness and height may play some role in humans' judgments, while the other heuristics do not. To evaluate the model without the effect of H_z and H_h, we computed the partial correlation between model and participants' responses while controlling out 1) height, which was 0.76 ± 0.02, and 2) center of mass ($z-$coordinate), which was 0.66 ± 0.03. This is consistent with the hypothesis that a physics model plays a larger, independent role in human judgments than heuristics like center of mass or height.

3.3.4 Experiment 2: "Direction" prediction

Methods

The second experiment asked participants to predict the direction that towers fall in. This experiment used a different set of tower stimuli which were all unstable. "Pairs" of two unstable mass configurations were again chosen for each geometric configuration and were selected to maximize the angle between fall directions within pairs. Unlike the "stability" experiment, we did not include a set of example towers prior to the training period. These examples did not appear to help participants in the stability experiment, and we believed there to be sufficient information about the ratio due to the fact that the towers were all unstable and feedback was given on every trial.

Participants were asked the question, "In what direction will this tower fall?". To indicate their response, participants were instructed to move a white line around the table to indicate the direction

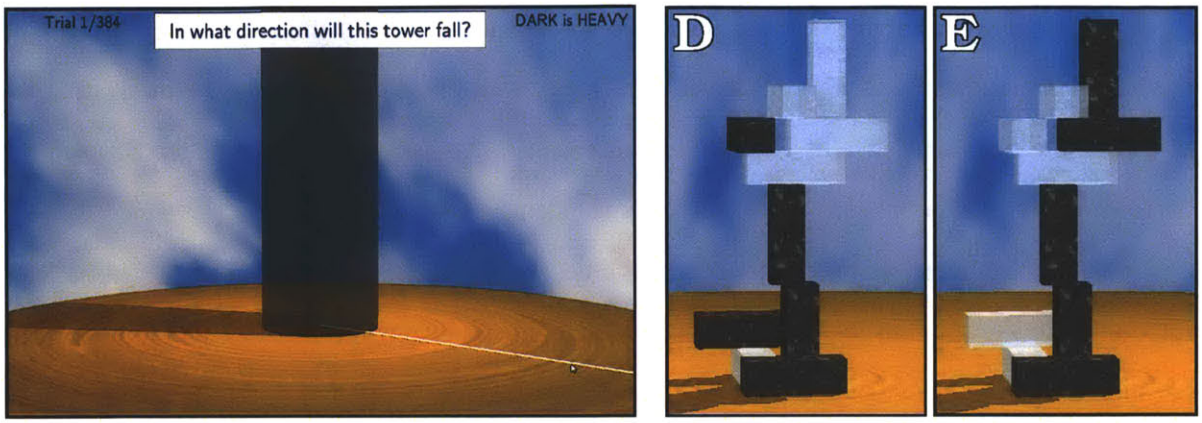


Figure 3.5: **Screenshot of the mass-sensitive "direction" prediction experiment.** Participants looked at stimuli such as Towers D and E, where the masses of the blocks (dark, heavy stone vs. light, translucent plastic) affected the direction that the tower would fall. They answered the question, "In what direction will this tower fall?" by adjusting the direction of a white line on the table, as in the screenshot on the left.

they thought the tower would fall (Figure 3.5).

Results

One of the key ideas of the rich simulation-based model is that it is able to easily generalize to many different tasks and situations. In order to assess this flexibility, we compared human judgments ($n = 10$), R_{dir}, regarding the direction a tower will fall, with the model's predictions, F_{dir}, as well as the skew heuristic prediction, H_{θ}.

Upper bound As in the stability results, we computed an upper bound on how well our model can do by correlating half of peoples judgments against the other half in a bootstrap analysis. This yielded a correlation of 0.73 ± 0.05, indicating a high degree of variability in peoples judgments.

Overall direction judgments The circular correlation between R_{dir} and F_{dir} was 0.48 ± 0.04 while the correlation between R_{dir} and H_{θ} was 0.09 ± 0.02. These correlations indicate that the model is far better at explaining humans' direction judgments and Figure 3.6 illustrates these results by plotting the differences between R_{dir} and F_{dir} for each tower (dots).

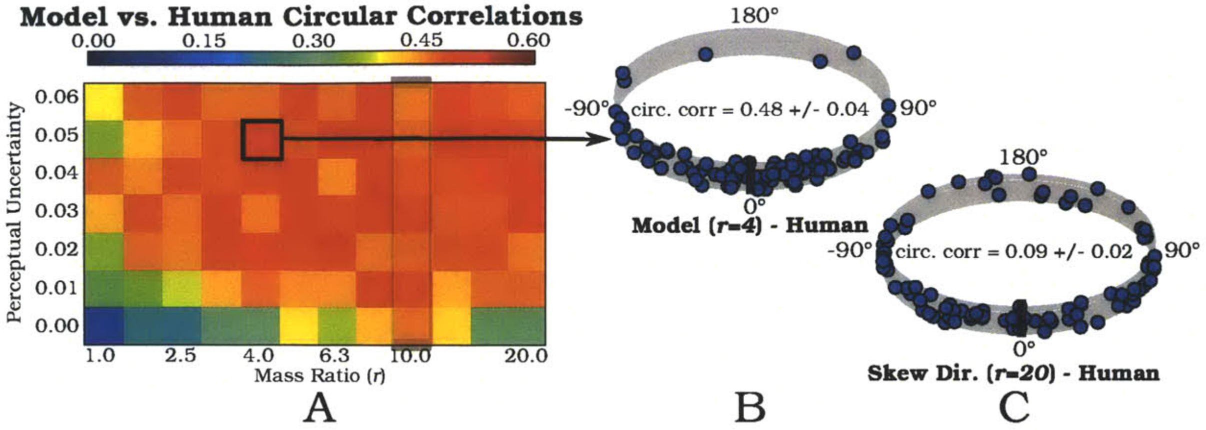


Figure 3.6: **Model and skew direction heuristic performance in the "direction" prediction experiment.** A shows circular correlations of model and human predictions, for different values of mass ratio (r) and perceptual uncertainty (σ_p). B and C show polar scatter plots of differences between model ($r = 4$) and human predictions (B) and the skew direction heuristic (H_{dir}) and human predictions.

Mass-sensitivity To determine whether people were making different judgments on different mass assignments, we computed differences between judgments for the two mass configurations for each geometric configuration. Correlating these differences for the model ($r = 4$) and people gives a circular correlation of 0.39 ± 0.06, whereas for the uniform-mass model ($r = 1$), the circular correlation is 0.11 ± 0.06. Thus, the mass-sensitive model is able to explain differences in people's judgments where the uniform-mass model cannot; this supports the hypothesis that people take the mass configurations into account when making their judgments.

Effect of variance The model's predictions about different towers' fall directions vary significantly in confidence, due to the effects perceptual uncertainty on different samples' physical outcomes. Confidence of fall direction judgments can be quantified by circular variance of the model's fall-direction estimates (insets in Figure 3.7) over the 20 simulations sampled from the same tower. In order to assess model fits on the stimuli for which model predictions are most meaningful, we sorted pairs by the circular variance of model predictions of the highest-variance mass assignment for each pair. We computed model-participant correlations for the k lowest-variance pairs, where k was varied from 5 to all 48 pairs. Figure 3.7 shows the circular correlations between model ($r = 4$)

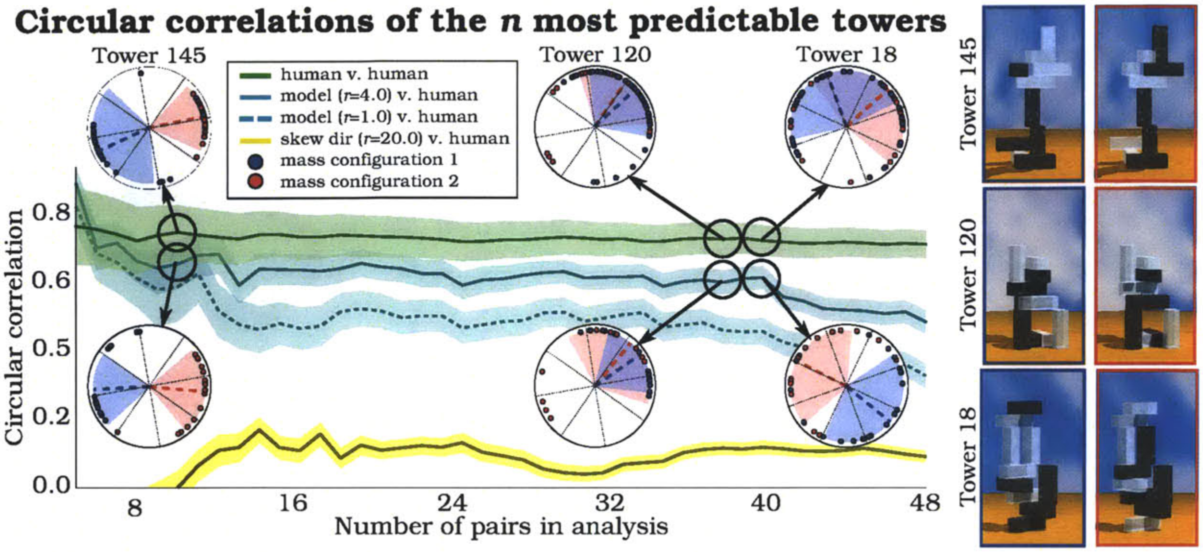


Figure 3.7: **Circular correlations as a function of model variance.** We sorted the pairs by the model's ($r = 4$) circular variance, comparing the highest variances of either mass configuration across pairs. We then computed circular correlations on the first k of these sorted pairs. This plot shows the correlation as a function of k for human v. human, model ($r = 4$ and $r = 1$) v. human, and skew direction (H_{dir}) v. human predictions. The polar insets show the distributions of human and model responses for three particular geometric configurations (screenshots of the mass configurations are shown on the right). In all cases, shaded regions show the standard error of the mean. Tower 145 shows clear differences in judgments for both the model and people for the two mass configuration. Neither the model nor people judge any difference between the mass configurations of Tower 120. There are some towers, like Tower 18, for which the model does not predict people well; however, these towers also tend to be the most variable.

and human predictions, as well as human vs. human predictions, the uniform-mass model vs. human predictions, and the skew direction heuristic (H_θ) vs. human predictions, as functions of k. Excluding the 8 pairs with lowest confidence (i.e., $k = 40$ or less) the correlation between R_{dir} and F_{dir} is reliably around 0.60. Predictions for direction of fall thus seem to match human judgments well in those cases where model predictions are meaningfully defined. However, people seem to predict each other at the same accuracy regardless of tower variance, suggesting that when towers are too difficult to predict, they fall back on another strategy that our model fails to capture.

3.4 Discussion

We report two important findings: first, that humans make different predictions of physical structural properties when the mass of objects may affect those properties. Second, that human physical reasoning is consistent with a model that uses internal knowledge of physical principles – including mass – to predict future scenes states, and that internal limitations like uncertainty due to noise can account for deviations from ground truth performance. These results are consistent with Hamrick et al. (2011) and provide further support for the simulation-based theory of physical reasoning.

While our model is a good predictor of human physical reasoning (Figures 3.4 and 3.6), people predict each others' responses even better. This suggests that there is systematic structure to people's judgments that our model does not capture. The model may be limited by its assumption that the brain perfectly models Newtonian dynamics, or its approximations of perceptual inference. One improvement might be to adopt noisy physics simulations, with accuracies diminishing rapidly over time. Another might be to vary perceptual uncertainty for blocks that are visible and occluded, respectively – which can be tested by manipulating participants' viewing conditions.

Although our simulation-based model is still preliminary, we emphasize its ability to generalize to many different tasks. We have shown that the same model can predict human judgments on stability and direction tasks, both in the context of uniform mass and when objects vary in mass. Alternate models, namely heuristic-based accounts, do not offer this flexibility from task to task. Indeed, it is frequently the case that even different versions of the *same* task need new heuristics: the best heuristic in the uniform-mass stability experiment, height, was not nearly as predictive as the center of mass in the mass-sensitive stability experiment.

Chapter 4

Physical Property Inference

4.1 Simulation-based inference model

Given an initial state and the ability to simulate forward, it is straightforward to predict the future. However, given a sequence of observations, it is not immediately clear how to infer the underlying properties. Here we present a novel model of human physical property inference that utilizes the same simulation-based reasoning as the prediction model from Chapter 3.

4.1.1 Motivation and overview

With the success of the simulation-based prediction model, we hypothesize that people may be using similar predictive mechanisms to perform inference of object properties. People performing a physical task (such as catching a ball) will make predictive eye movements to future collision points as they watch a dynamic event unfold (Hayhoe, Mennie, Sullivan, & Gorgos, 2005; Zago, McIntyre, Senot, & Lacquaniti, 2009). Relevantly, Hayhoe et al. (2005) report that such predictive eye movements from observers in a ball-catching task may reflect updates to an internal model of the ball's dynamics in response to errors. While there are fewer indications of this behavior from observers who are merely watching the ball being caught, Hayhoe et al. (2005) suggest that this may be because non-catchers are not as invested in maintaining the tracking precision required to catch a ball.

We propose the hypothesis that people can infer object properties (e.g., mass) through a process of prediction, observation, and belief updating in response to discrepancies between predictions and observations. This idea is consistent with the work of Sanborn et al. (2009), who used a similar probabilistic model to infer object mass ratios after observing a single collision. However, we are interested in scenarios where many collisions may occur (possibly even at the same time).

Thus, we focus on a model that integrates information across both multiple points in time and multiple objects by updating its beliefs in response to prediction errors.

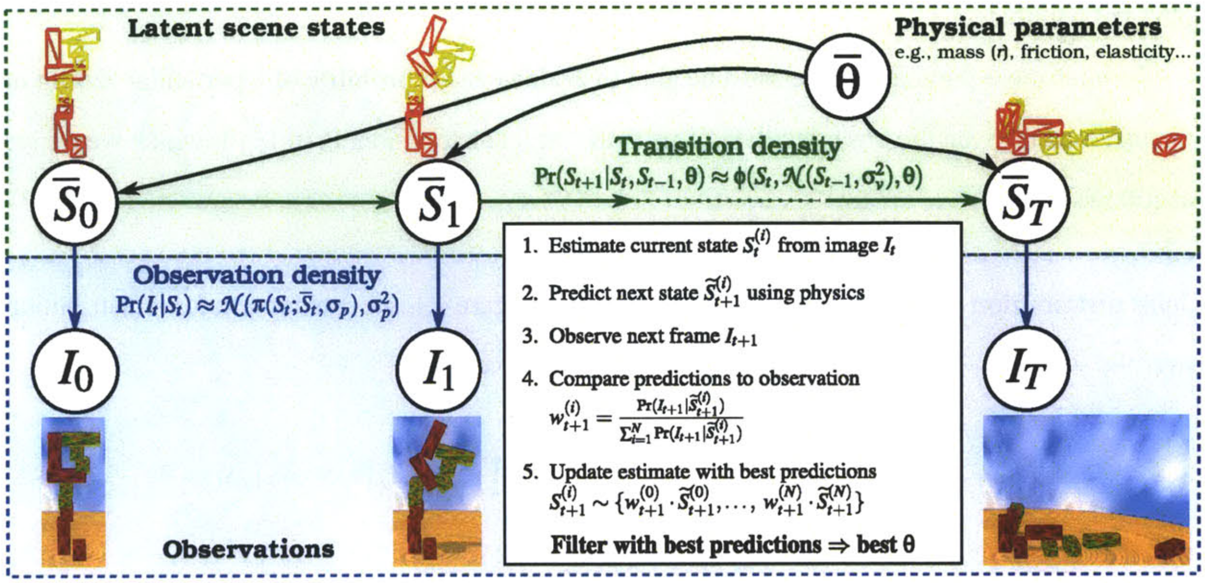
This strategy of belief updating aligns well with a sequential time-series Monte Carlo inference technique from machine learning called *particle filtering* (Cappé, Godsill, & Moulines, 2007). Particle filters have been hugely successful in many computer vision motion tracking applications (Ristic, Arulampalam, & Gordon, 2004), including those involving multiple interacting objects (Khan, Balch, & Dellaert, 2004). Generally speaking, particle filters operate in the following manner:

1. Sample a distribution of latent scene states given an observation (see "Scene samples" in Section 3.1). Each sample is referred to as a "particle".
2. Predict the dynamics for the next time delta for each of these particles (see "Physics simulator" in Section 3.1).
3. Compute weights for each particle based on how closely it matches the new observation.
4. Resample with replacement a new set of particles from the old weighted particles.
5. Repeat from step 2.

The weighted particles at each time step approximate a distribution over true scene states. In the following sections, we demonstrate how this distribution may be used to infer underlying physical parameters such as mass.

4.1.2 Definition

Figure 4.1 shows the causal structure of a physical event F as a standard Hidden Markov Model (Baum & Petrie, 1966). The observed event consists of a sequence of observations, $\mathbf{I}_{0:T} = \{I_0, \ldots, I_T\}$, where T is the number of observed frames.


Latent scene states
Physical parameters
e.g., mass (r), friction, elasticity...
$\overline{\theta}$
Transition density
$\Pr(S_{t+1}|S_t, S_{t-1}, \theta) \approx \phi(S_t, \mathcal{N}((S_{t-1}, \sigma_v^2), \theta)$
$\overline{S}_0$
$\overline{S}_1$
$\overline{S}_T$
Observation density
$\Pr(I_t|S_t) \approx \mathcal{N}(\pi(S_t; \overline{S}_t, \sigma_p), \sigma_p^2)$
I_0
I_1
I_T
Observations

1. Estimate current state $S_t^{(i)}$ from image I_t
2. Predict next state $\widetilde{S}_{t+1}^{(i)}$ using physics
3. Observe next frame I_{t+1}
4. Compare predictions to observation $w_{t+1}^{(i)} = \frac{\Pr(I_{t+1}|\widetilde{S}_{t+1}^{(i)})}{\Sigma_{i=1}^{N} \Pr(I_{t+1}|\widetilde{S}_{t+1}^{(i)})}$
5. Update estimate with best predictions $S_{t+1}^{(i)} \sim \{w_{t+1}^{(0)} \cdot \widetilde{S}_{t+1}^{(0)}, \ldots, w_{t+1}^{(N)} \cdot \widetilde{S}_{t+1}^{(N)}\}$

Filter with best predictions ⇒ best θ

Figure 4.1: **Physical parameter inference model.** 1. People first observe the scene and form beliefs, $S_t^{(i)}$, about the true state of the scene, $\overline{S}_t$. These beliefs incorporate some amount of perceptual uncertainty, quantified by σ_p. 2. They then predict possible next states, $\widetilde{S}_{t+1}^{(i)}$ using a process of noisy physical reasoning approximating Newtonian physics. The uncertainty in this process is quantified by σ_d. 3. After some amount of time, people observe the scene again. 4. They compare their predictions to their observation and weigh the predictions by their accuracy. 5. Finally, they update their estimates of the latent scene state based on which predictions were most accurate.

Scene states The estimated scene states S_t are given by the position and quaternion vectors of all objects in the scene. As in Section 3.1, perceptual beliefs about these states are modeled as posterior inference, where the scale of objects' positional uncertainty is quantified by the σ_p parameter. Again, we use $\mathbf{S}_{0:T}$ as the sequence of states from $t = 0$ to $t = T$.

Physical dynamics The process of physical dynamics imposes a dependency of each scene state on the state before it and on the underlying physical parameters of each object in the scene. In Section 3.1, we used $S_{t+T} = \Phi(S_t, T, r)$ to denote physical dynamics with respect to mass ratio. For the inference model, we break Φ into individual time steps, $S_{t+1} = \phi(S_t, S_{t-1}, r)$.

Physical parameters Previously, we used r to represent the mass ratio. However, the inference model may apply to any physical parameter estimation; thus, we will use $\overline{\theta}$ to denote the entire set of *true* physical parameters (e.g., mass, friction, elasticity) and θ to denote the inferred set of physical parameters.

To infer these parameters, we must be able to evaluate the probability of a particular setting of the parameters given the observed data. This is the target distribution, $\Pr(\theta|\mathbf{I}_{0:T})$, which we derive as follows.

Joint distribution From the Markov chain shown in Figure 4.1, we can define a joint distribution over the states $\mathbf{S}_{0:T}$, observations $\mathbf{I}_{0:T}$, and physical parameters θ:

$$\Pr(\mathbf{I}_{0:T},\mathbf{S}_{0:T},\theta) = \Pr(\theta)\Pr(I_0|S_0)\Pr(S_0|\theta)\cdot\prod_{t=1}^{T}\Pr(I_t|S_t)\Pr(S_t|S_{t-1},S_{t-2},\theta) \tag{4.1}$$

This distribution contains two key parts: the *observation density* and the *transition density*. The observation density $\Pr(I_t|S_t,\sigma_p)$ is the probability that the observed data I_t was generated by a given state S_t. The transition density $\Pr(S_t|S_{t-1},S_{t-2},\theta)$ is the probability of reaching state S_t given the previous state and the parameters. Because $\phi(S_{t-1},S_{t-2},\theta)$ is deterministic, however, this becomes a delta function. To avoid this problem, we assume some amount of uncertainty in how the dynamics – specifically, velocity – are computed. We introduce the σ_d parameter, which quantifies this dynamics uncertainty.

Target distribution To determine the *target distribution* $\Pr(\theta|\mathbf{I}_{0:T})$, we must first compute $\Pr(\mathbf{I}_{0:T},\theta)$ in terms of the joint distribution (Equation 4.1), the observation density, and the transition density:

$$\Pr(\mathbf{I}_{0:T},\theta) = \Pr(\theta)\int_{\mathbf{S}_{0:T}}\Pr(I_0|S_0)\Pr(S_0|\theta)\cdot\prod_{t=1}^{T}\underbrace{\Pr(I_t|S_t)}_{\substack{\text{observation}\\\text{density}}}\underbrace{\Pr(S_t|S_{t-1},S_{t-2},\theta)}_{\substack{\text{transition}\\\text{density}}}\,\mathrm{d}\mathbf{S}_{0:T} \tag{4.2}$$

To obtain the target distribution itself, we normalize (4.2) by its integral with respect to θ:

$$\Pr(\theta|\mathbf{I}_{0:T}) = \frac{\Pr(\mathbf{I}_{0:T},\theta)}{\int_\theta \Pr(\mathbf{I}_{0:T},\theta)\,d\theta} \tag{4.3}$$

Model predictions For simplicity, we restrict our model to cases in which there are two types of objects. We wish to infer the mass ratio between these object types; thus, as its judgment, $\overline{F}_{light}$, our model computes the probability that the mass ratio is less than 1:

$$\overline{F}_{light} = \Pr(\theta < 1|\mathbf{I}_{0:T}) = \int_0^1 \Pr(\theta|\mathbf{I}_{0:T})\, d\theta \tag{4.4}$$

4.1.3 Algorithm

To compute the integral in (4.2), we use a particle filter with particle rejuvenation (Cappé et al., 2007). In the general case, particle filters allow us to approximate the distribution $\Pr(S_t|I_t)$ via a set of samples $\{S_t{}^{(0)}, \ldots,\ S_t{}^{(N)}\}$. Pseudocode for this algorithm can be found in Algorithm 4.2.

Observation density We model the uncertainty in the current state of the object in the same manner as the prediction model in Section 3.1:

$$\Pr(I_t|S_t) \approx \mathcal{N}(\pi(S_t;\ \overline{S}_t, \sigma_p),\ \sigma_p{}^2) \tag{4.5}$$

Similar to the π of the prediction model, $\pi(\cdot)$ here is a set of independent Gaussian distributions for each objects' x, y and z positions, with variance $\sigma_p{}^2$ and mean centered on the corresponding block positions in the true world state $\overline{S}_0$, followed by a deterministic transform that prevents blocks from interpenetrating.

Transition density We also model dynamics uncertainty using a Gaussian distribution over the velocity produced by the deterministic $\phi(\cdot)$:

$$\Pr(S_t|S_{t-1}, S_{t-2}, \theta) \approx \phi(S_{t-1},\ \mathcal{N}(S_{t-2}, \sigma_d{}^2),\ \theta) \tag{4.6}$$

Filtering To obtain a distribution over states for each time step, we iteratively perform the following computations:

Propagate particles A new set of samples are drawn from the *proposal distribution* (4.7) and weights are computed for each sample (4.8):

$$\widetilde{S}_t^{(n)} \sim \Pr(S_t | S_{t-1}{}^{(n)},\ S_{t-2}{}^{(n)},\ \theta) \tag{4.7}$$

$$\widetilde{w}_t^{(n)} = w_{t-1}^{(n)} \cdot \frac{\Pr(I_t | \widetilde{S}_t^{(n)}) \cdot \Pr(\widetilde{S}_t^{(n)} | S_{t-1}{}^{(n)},\ S_{t-2}{}^{(n)},\ \theta)}{\Pr(\widetilde{S}_t^{(n)} | S_{t-1}{}^{(n)},\ S_{t-2}{}^{(n)})} \tag{4.8}$$

In the general case, the proposal distribution approximates the transition density (which cannot be directly sampled from, but may be evaluated), and so the weights must be scaled to reflect the difference between the two distributions. In our case, however, the proposal distribution is equivalent to the transition density. Thus, (4.8) becomes:

$$\widetilde{w}_t^{(n)} = w_{t-1}^{(n)} \cdot \Pr(I_t | \widetilde{S}_t^{(n)}) \tag{4.9}$$

Normalize weights We now normalize the weights such that they sum to 1:

$$w_t^{(n)} = \frac{\widetilde{w}_t^{(n)}}{\sum_{i=1}^{N} \widetilde{w}_t^{(i)}} \tag{4.10}$$

Resample particles Finally, we resample the particles based on these weights, such that their distribution reflects the posterior for that timestep, and uniformly reset the weights (because the particles now implicitly represent the distribution, we no longer require the weights to provide that information):

$$S_t{}^{(n)} \sim \Pr(S_t | \mathbf{I}_{0:t}, \theta) \propto \Pr(I_t | \widetilde{S}_t^{(n)}) \cdot \Pr(\widetilde{S}_t^{(n)} | S_{t-1}{}^{(n)},\ S_{t-2}{}^{(n)},\ \theta) \tag{4.11}$$

$$w_t^{(n)} = \frac{1}{N} \tag{4.12}$$

Computing the likelihood We are finally ready to compute the likelihood of the data given the parameters. The normalization in (4.10) is very important, as we will use the denominator to find $\Pr(\mathbf{I}_{0:T}, \theta)$. Here we examine how this relates to (4.2). First, the weights have values equal to

$\Pr(I_t|\widetilde{S}_t^{(n)})$, where:

$$\widetilde{S}_t^{(n)} \sim \Pr(S_t|S_{t-1}{}^{(n)},\ S_{t-2}{}^{(n)},\ \theta)\Pr(S_{t-1}{}^{(n)}|\mathbf{I}_{0:t-1},\theta) \propto \Pr(S_t|\mathbf{I}_{0:t-1},\theta) \tag{4.13}$$

Because we know that the information from all of the previous observations is contained in the samples, and the weights are computed from the samples, we can rewrite the denominator of (4.10):

$$\sum_{i=1}^{N} \widetilde{w}_t^{(i)} = \sum_{i=1}^{N} \Pr(I_t|\widetilde{S}_t^{(i)}) = \Pr(I_t|\mathbf{I}_{0:t-1},\theta) \tag{4.14}$$

If we now take the product of these values across time, we obtain the likelihood of the data:

$$\prod_{t=1}^{T} \Pr(I_t|\mathbf{I}_{0:t-1},\theta) = \Pr(\mathbf{I}_{0:T}|\theta) \tag{4.15}$$

Which is equivalent to the integral over $\mathbf{S}_{0:T}$ in (4.2).

Model predictions As before, to compute the model judgment, F_{light}, we compute the probability that the mass ratio between two different types of objects is less than 1. We approximate the integral in (4.4) through a set of parameter samples:

$$\overline{F}_{light} \approx F_{light} = \sum_{\theta<1} \left(\frac{\Pr(\mathbf{I}_{0:T}|\theta)\Pr(\theta)}{\sum_{\theta} \Pr(\mathbf{I}_{0:T}|\theta)} \right) \tag{4.16}$$

```
Input: $\mathbf{I}_{0:T}, \theta, \Pr(\theta)$
 1: // Initialize
 2: for $n = 1 \rightarrow N$ do
 3:     Sample $S_0^{(n)} \sim \Pr(S_0|I_0)$
 4:     $w_0^{(n)} \leftarrow \frac{1}{N}$
 5: end for

 6: for $t = 1 \rightarrow T$ do
 7:     // Propagate particles
 8:     for $n = 1 \rightarrow N$ do
 9:         Sample $\widetilde{S}_t^{(n)} \sim \Pr(\widetilde{S}_t^{(n)}|S_{t-1}^{(n)}, \theta)$
10:         $\widetilde{w}_t^{(n)} \leftarrow \Pr(I_t|\widetilde{S}_t^{(n)})$
11:     end for

12:     // Normalize weights
13:     Let $\Pr(I_t|\theta) \approx \sum_n \widetilde{w}_t^{(n)}$
14:     $w_t^{(n)} = \Pr(I_t|\theta)^{-1} \cdot \widetilde{w}_t^{(n)}$ for $n \in \{1, \ldots, N\}$

15:     // Resample particles
16:     for $n = 1 \rightarrow N$ do
17:         Let $\Pr(n) \approx w_t^{(n)}$
18:         Sample $n_j \sim \Pr(n_j)$
19:         $S_t^{(n)} \leftarrow \widetilde{S}_t^{(n_j)}$
20:     end for
21: end for

22: return $\{\{(\widetilde{S}_t^{(i)}, S_t^{(i)}, \widetilde{w}_t^{(i)}, w_t^{(i)})\}_{i=1}^{N}\}_{t=0}^{T}$
```

Figure 4.2: **Particle filter-based inference algorithm**. Lines 1-5: Initialize the particle filter with a set of scene samples based on observing the image. Lines 7-11: Using non-deterministic physical dynamics approximating Newtonian physics, predict the state of the particles on the next time step, observe the new scene, and weigh the particles based on how close they approximate the observation. Lines 12-14: Normalize the weights such that they sum to 1. Lines 15-20: Resample the particles with replacement based on the normalized weights. Line 22: Return the full set of auxiliary scene states, resampled scene states, (un-normalized) auxiliary weights, and normalized weights for every time step.

4.2 Visual and geometric heuristics for inference

Previous work on mass inference suggests that people may be relying on a variety of heuristics rather than a simulation-based model (Todd & Warren, 1982; Gilden & Proffitt, 1989a, 1989b, 1994). We examine several such heuristics pertaining to object velocity; for each of the following heuristics, the objects with higher values are chosen to be the lighter objects:

- $H_{\max(v)}$ (maximum velocity): the maximum velocity over all time steps
- H_d (maximum movement): the maximum distance traveled over all time steps. Note that this is *not* displacement; it is the length of the whole path that the object travels.
- $H_{\mathrm{total}(d)}$ (total movement): the total distance traveled by objects with the same mass.
- $H_{\mathrm{avg}(d)}$ (average movement): the average distance traveled by objects with the same mass.

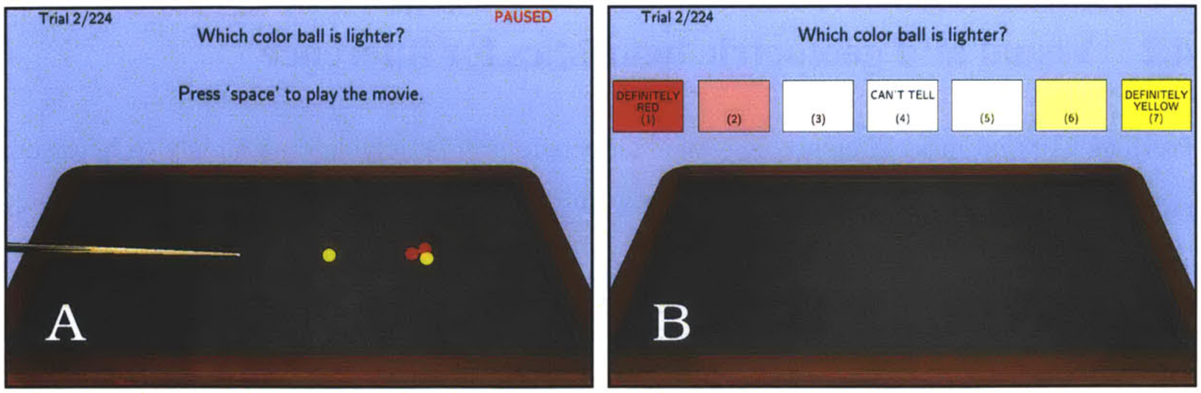
4.3 Experiment 3: "Billiards inference" (collisions)

Previous mass inference experiments showed participants 2D scenes, usually of only two objects colliding. However, the prediction experiments from Chapter 3 indicated that people can reason about physical events when there are objects of different masses. Thus, we also look at more realistic scenarios in the context of mass inference.

4.3.1 Methods

Participants Another 4 adult participants were recruited from the MIT BCS human subject pool. As in the prediction experiments, each gave informed consent in accordance with MIT's IRB. They were compensated $10/hour, and performed 1 hour each.

Apparatus Participants viewed trials on the same computer as in the prediction experiments and simulations were again rendered at 1000Hz. As before, all stimuli were rendered in 3D using



Trial 2/224

Which color ball is lighter?

PAUSED

Press 'space' to play the movie.

Trial 2/224

Which color ball is lighter?

Figure 4.3: **"Billiards inference" experiment screenshot.** Participants watched movies of billiards-like scenes, such as in A. The pool cue stick would move forward, hitting the cue ball, which in turn would collide with the rest of the balls. The yellow balls always had the same mass, m_{yellow}, and the red balls always had the same mass m_{red}, but the ratio of these masses, $r = \frac{m_{red}}{m_{yellow}}$, could be any of 1:10, 1:5, 1:2, 1:1, 2:1, 5:1, or 10:1. After watching the movie, participants judged which color was lighter on a graded scale of confidence from "definitely red" (1) to "definitely yellow" (7), as shown in B.

Panda3D (CMU Entertainment Technology Center, 2012) and physics simulations were computed at 1000Hz using the ODE physics engine (Smith, 2009). Animations of the physical dynamics were played back at 1x speed. Participants submitted response judgments by depressing keyboard keys.

Trials depicted a 3D scene that contained a 9.777m x 5.437m felt-covered rectangular table with raised edges. Unlike a real pool table, this table did not have pockets for the balls (Figure 4.3). A "pool cue" stick with length 5.25m and radius 9cm was placed on the left of the scene, hovering above the table and slightly angled down at 2.5° such that the tip was 12.8cm above the table. Several balls with radius 12.8cm were arranged in configurations to the right of the pool cue. The pool cue had a coefficient of elasticity of $e = 0.4$, while all balls were perfectly elastic ($e = 1.0$).

Stimuli Stimuli were created from 3 basic configurations in which a "cue ball" was placed 1.625m to the right of the pool cue and a group of "racked balls" was placed 1.539m to the left of the cue ball (Figure 4.4). For each basic configuration, some fraction of the balls were chosen to be

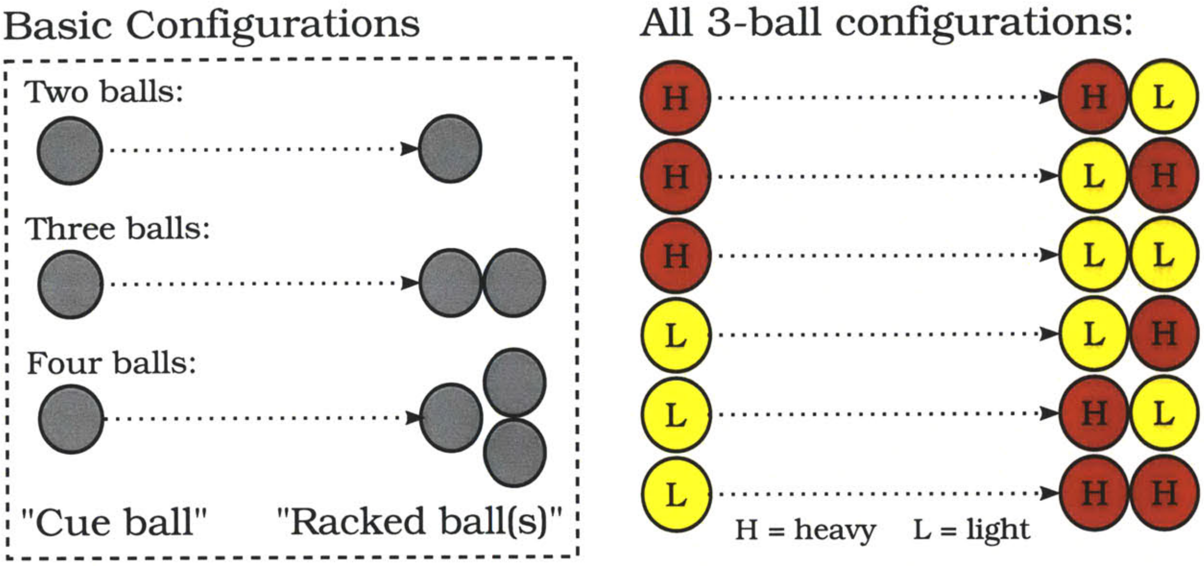


Figure 4.4: **"Billiards inference" stimuli configurations.** On the left are the three "basic configurations" of two, three, and four balls. The ball on the left was always the "cue ball" and during physical simulation would be hit by the cue stick. The ball(s) on the right were always the "racked balls" and were impacted by the cue ball. On the right are all possible heavy/light configurations for three balls. Note that while red is heavy in this diagram, the colors were counterbalanced during the experiment.

heavy or light and we created configurations for every possible permutation of heavy/light balls. For each of these permutations, we assigned the masses of the balls according to several different heavy-to-light mass ratios: 1:1, 2:1, 5:1, and 10:1. Finally, for each of these stimuli, the "racked balls" were moved a small amount along the $y-$axis, where the amount of shift was drawn from a zero-mean Gaussian distribution with a standard deviation of 1.7cm. This was to ensure that the cue ball did not hit the racked balls perfectly head-on.[1]

Trial structure All trials had 3 phases: *stimulus-snapshot*, *stimulus-dynamics*, and *response*.

Stimulus-snapshot phase. The *stimulus-snapshot* phase began with a static presentation of the stimulus. The camera radius was 12m and the field of view was 40°. Participants were allowed to observe this static image for as long as they wished, and moved to the next phase by pressing the

[1] When the cue ball and racked balls are perfectly aligned, the resulting collisions look bizarre and "too perfect". The $y-$axis noise creates scenarios that are more realistic, as we rarely see perfectly-aligned collisions in the real world.

space bar. A screenshot of this phase is shown in Figure 4.3A.

Stimulus-dynamics phase. The *stimulus-dynamics* phase began immediately after the stimulus-snapshot phase and lasted 3500ms. During this phase, physics was turned "on" and the pool cue moved forward at $9\frac{m}{s}$, impacting the "cue ball". This caused the cue ball to move forward at a velocity of $70\frac{m}{s}$, colliding with the other balls in the scene. To simulate the rolling friction of the felted table, the balls' velocities were damped at a rate of 0.9984, i.e., the balls lost 0.16% of their linear and angular velocities every millisecond. The stimulus was removed from the scene after 3500ms and the response phase began.

Response phase. The *response* phase was not limited in time, but ended once the participant depressed a response key. Afterward, there was a delay of 1000ms before the next trial's stimulus-snapshot phase began.

On each trial participants made graded responses to the question, "Which color ball is lighter?", by pressing keys on a 1-7 scale to indicate degrees of confidence between "definitely red" (1) to "definitely yellow" (7) (Figure 4.3B). Participants judged 4 subsessions of 56 trials (224 total trials), where each subsession consisted of 2- or 4-ball stimuli (as described in the previous section). Over the course of the experiment, these stimuli were repeated four times (with counterbalanced colors) in a random order, such that there were at least 2 trials between repetitions of the same stimulus. Participants were not told that stimuli were repeated.

Before the actual experiment, participants judged 18 "training" trials to be familiarized with the task. These trials consisted of only 3-ball stimuli with heavy-to-light mass ratios of 10:1 or 1:1. Otherwise, the format of the trials was identical to the actual experiment.

4.3.2 Analysis

On each trial, we present a billiards-like scene with state sequence $\overline{\mathbf{S}}_{0:T}$ and the human observer responds with the lighter color, R_{light}. We computed the mean R_{light} across participants for each scene as their collective mass inference. Similarly, we computed the mean across model samples,

F_{light}, for each scene as its mass inference. We performed correlation analyses using Pearson's correlations and these correlations were computed using a bootstrap analysis (repeated 1000 times, sampling with replacement) over individual human and model judgments. We computed correlations between mean human and model judgments for each bootstrap sample, and from these computed the mean correlations as well as standard error.

4.3.3 Results

We compared human judgments ($n = 4$) to the inference model's judgments as well as to several heuristics. The correlation between people and the model was 0.61 ± 0.02, whereas the maximum movement heuristic H_d gave a correlation of 0.86 ± 0.02 with people. Several other heuristics also explained people's judgments better than the model: total movement, $H_{\mathrm{total}(d)}$, had a correlation of 0.69 ± 0.02; and average movement, $H_{\mathrm{avg}(d)}$, had a correlation of 0.81 ± 0.02. The maximum velocity heuristic, $H_{\mathrm{max}(v)}$, did not explain people's judgments well, with a correlation of 0.40 ± 0.03.

These results suggest that in this task, people are not using a simulation-based approach to property inference, or at least, not relying solely on it. It is likely that they are instead relying primarily on simple visual heuristics – in particular, how much objects move – to judge which color object is heavier.

4.4 Discussion

We proposed a novel model of human physical property inference based on the notion of updating beliefs about those properties in response to prediction errors. However, preliminary results from a mass-inference task do not support the hypothesis that people use such a model: rather, it was likely that people were using simple perceptual cues based on velocity and distance. These results are consistent with a "perceptual heuristics" theory (Todd & Warren, 1982; Gilden & Proffitt, 1989a, 1989b, 1994).

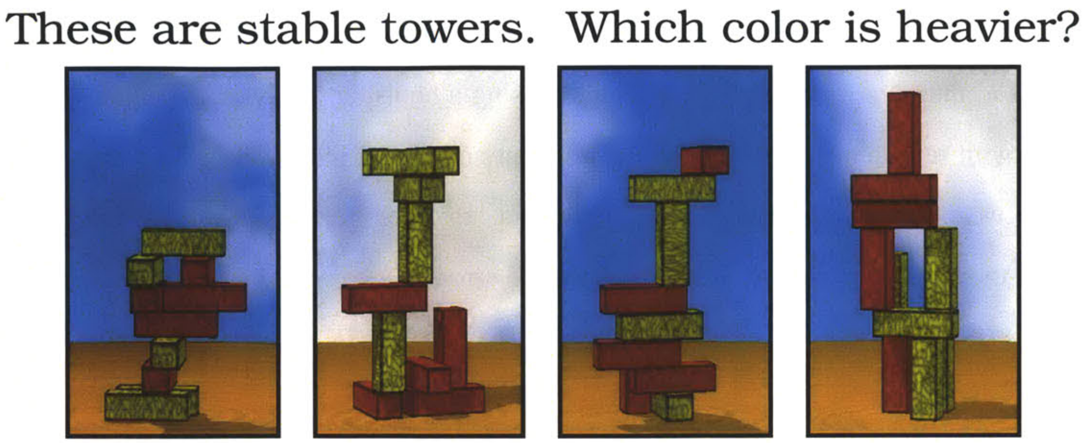


Figure 4.5: **Mass inference task for stable towers.** As in Figure 2.2, all of the red blocks have the same mass m_{red} and all of the yellow blocks have the same mass m_{yellow}. One of the colors *must* be heavier for these towers to be stable. For each tower, which color is heavier? (Answer: yellow, red, yellow, yellow)

Yet, there is evidence that people use a richer model in prediction-based tasks (see Chapter 3). Several possibilities for why people did not use a similar approach in the billiards experiment present themselves. First, it is possible that the task was too hard. The events in the movies happened very quickly and people may not have been able to process the information from collisions fast enough to use a simulation-based model. Moreover, both participants and colleagues have reported that they would have liked to see the movies more than once, suggesting that people have difficulty in remember the initial scene state. A second possibility is that heuristics are just generally more reliable. Indeed, billiards is not as simple and deterministic as it might seem; after about 9 collisions, the players' gravity will begin to have a non-negligible impact on the dynamics of the balls (Berry, 1978). Perhaps people pick up on this unpredictability and resort to the heuristics, which, while not 100% accurate, are likely sufficient.

Despite the results of this experiment, it is still possible that people use a simulation-based approach to physical property inference in certain scenarios. A pilot study of the "stable inference task" (Figure 4.5) suggests that naïve subjects have trouble with the inference task. However, colleagues familiar with how objects behave in simulation find the task straightforward and natural. It is possible that participants do not quite know what to expect of these "video game"-like scenarios, and first need more exposure to the dynamics. To test this idea, we plan to re-recruit participants from the stability and direction experiments in the previous chapter to perform this task.

Another future approach is to recruit participants for the stability and direction tasks, but replace the stone and plastic blocks with red and yellow blocks, as in Figure 4.5. Participants would then determine which color was heavier over the course of the experiment. As we have previously demonstrated, participants appear to use a simulation-based approach in the stability direction tasks; thus, it is likely that this modified experiment would present an ideal scenario for simulation-based inference as well.

Chapter 5

Conclusion

5.1 Contributions

We present three contributions to the fields of cognitive science and computer science.

First, this work continues to support the hypothesis that knowledge of Newtonian principles and probabilistic representations are generally applied for human physical reasoning in prediction-based scenarios. Complex tasks like predicting the stability of a tower of blocks are both expressible in our modeling framework and well-matched with human performance. This idea provides rich and flexible foundational groundwork for developing a comprehensive model that naturally scales to a broad class of human physical judgments. Moreover, this model is conducive to artificial intelligence and robotics applications. Many handheld and mobile devices (e.g., iPhones) have physics-based games that run physical simulations in real time. We also structure our model around a video-game physics engine; thus, it should be fairly transferable to portable and possibly even embedded devices.

Second, we demonstrate that people are able to reason about physical properties such as mass in complex, chaotic scenarios. This may be surprising in light of previous work on intuitive physics with much simpler situations focusing on the ways in which human judgments are biased and error-prone (Todd & Warren, 1982; Gilden & Proffitt, 1989a, 1989b, 1994; Caramazza et al., 1981; McCloskey, 1983). Future work will explore the differences between our tasks and previous intuitive physics studies that might explain this gap, such as differences in the ecological validity of the scenarios, stimuli and tasks (Zago & Lacquaniti, 2005).

Third, we proposed a novel framework of human property inference. Although we were unable to demonstrate that people use something other than simple, perceptual heuristics in a billiards-like mass inference task, we cannot rule out the possibility that there are other scenarios more conducive to a simulation-based approach. Regardless of its applicability to human reasoning, this inference model has applications to robotics and artificial intelligence, allowing a robot to fluently reason about the physical properties of objects in the world around it.

5.2 Future work

Future work will, in particular, focus on the capabilities of humans to perform inference of unobservable physical properties and attempt to find situations in which people are more likely to approach the problem through simulation-based reasoning. Additionally, the reliance on heuristics in the billiards experiment brings up an interesting point about how people choose between a rich, simulation-based model vs. heuristics on certain tasks. It is likely that people use both simulation and heuristics simultaneously when both provide useful information (e.g., the height heuristic in Hamrick et al. (2011)). How are the proportions of this combination determined? Moreover, how are the heuristics learned?

A second focus of future work will be the inference model itself. In its current incarnation, it has difficulty making predictions of more than a few objects: for example, a tower of 10 blocks proves to be highly chaotic and difficult for the model to track. A better model might track only a handful of objects, allowing it to spend more of its computation time on improving accuracy for those few objects. Additionally, we made the decision to not update parameter estimates *during* particle filter runs because in reality, the parameters are fixed. However, it is possible that updating the estimates in real-time may actually be more representative of humans' reasoning process.

In sum, we find that rich internal physics models are likely to play a key role in guiding human common-sense reasoning in prediction-based tasks, including those where objects may be of variable mass. Inferring object mass is much more difficult and is possibly limited to a reliance on shallow heuristic models. However, we believe it more likely that people use a combination of simulation and heuristics to perform these inferences; we leave it as a direction for future work to determine the nature of this relationship.

Bibliography

Arimoto, S. (1999). Robotics research toward explication of everyday physics. *International Journal of Robotics Research*, *18*(11), 1056–1063.

Baillargeon, R. (1987). Young infants' reasoning about the physical and spatial properties of a hidden object. *Cognitive Development*, *2*(3), 179–200.

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

Baillargeon, R. (2007). The acquisition of physical knowledge in infancy: A summary in 8 lessons. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development.* Blackwell.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, *37*, 1554–1563.

Berry, M. V. (1978). Regular and irregular motion. *Proceedings of the Conference of the American Institute of Physics*, *46*, 16–120.

Cappé, O., Godsill, S. J., & Moulines, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, *95*(5), 899–924.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in 'sophisticated' subjects: misconceptions about trajectories of objects. *Cognition*, *9*(2), 117–123.

CMU Entertainment Technology Center. (2012). *Panda3D: Free 3D game engine.* Available from `http://www.panda3d.org/`

Fleming, R. W., Barnett-Cowan, M., & Bülthoff, H. H. (2010). Perceived object stability is affected by the internal representation of gravity. *Perception*, *39*, 109.

Forbus, K. D. (1981). *A study of qualitative and geometric knowledge in reasoning about motion.* M.S. thesis, Massachusetts Institute of Technology.

Forbus, K. D. (1983). Qualitative reasoning about space and motion. In D. Gentner & A. Stevens (Eds.), *Mental models.* New Jersey: LEA Associates, Inc.

Gardin, F., & Meltzer, B. (1989). Analogical representations of naive physics. *Artificial Intelligence*, *38*(2), 139–159.

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

Gilden, D. L., & Proffitt, D. R. (1989b). Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 384–393.

Gilden, D. L., & Proffitt, D. R. (1994). Heuristic judgement of mass ratio in two-body collisions. *Perception and Psychophysics*, *56*(6), 708–720.

Hamrick, J., Battaglia, P., & Tenenbaum, J. (2011). Internal physics models guide probabilistic judgments about object dynamics. *Proceedings of the 33rd Conference of the Cognitive Science Society.*

Hayes, P. (1978). The naive physics manifesto.

Hayes, P. (1984). The second naive physics manifesto. *Ubiquity.*

Hayhoe, M., Mennie, N., Sullivan, B., & Gorgos, K. (2005). The role of internal models and prediction in catching balls. *Proceedings of AAAI.*

Hecht, H. (1996). Heuristics and invariants in dynamic event perception: Immunized concepts or non-statements? *Psychonomic Bulletin and Review*, *3*, 61–70.

Jacobs, D. M., Michaels, C. F., & Runeson, S. (2000). Learning to perceive the relative mass of colliding balls: the effects of ratio scaling and feedback. *Perception and Psychophysics*, *62*(7), 1332–1340.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *The Journal of Neuroscience*, *27*(45), 12176–12198.

Khan, Z., Balch, T., & Dellaert, F. (2004). An MCMC-based particle filter for tracking multiple interacting targets. *Proceedings of the European Conference on Computer Vision*, *3024*, 279-290.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434–1448.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–130.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, *210*(4474), 1139–1141.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Ristic, B., Arulampalam, S., & Gordon, N. (2004). *Beyond the Kalman filter: Particle filters for tracking applications*. Artech House.

Runeson, S. (1977). *On visual perception of dynamic events*. S. Runeson.

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

Sanborn, A., Mansinghka, V., & Griffiths, T. (2009). A Bayesian framework for modeling intuitive dynamics. *Proceedings of the 31st Conference of the Cognitive Science Society*.

Smith, R. (2009). *Open dynamics engine*. Available from `http://www.ode.org/`

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

Tenenbaum, J., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

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

Toussaint, M., Plath, N., Lang, T., & Jetchev, N. (2010). Integrated motor control, planning, grasping and high-level reasoning in a blocks world using probabilistic inference. *Proceedings of the 2010 IEEE International Conference on Robotics and Automation (ICRA)*, 385–391.

Wolpert, D., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*(7-8), 1317–1329.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zago, M., & Lacquaniti, F. (2005). Visual perception and interception of falling objects: a review of evidence for an internal model of gravity. *Journal of Neural Engineering*, *2*, 198.

Zago, M., McIntyre, J., Senot, P., & Lacquaniti, F. (2009). Visuo-motor coordination and internal models for object interception. *Experimental Brain Research*, *192*, 571-604.